

STRUCTURE IN THE 3D GALAXY DISTRIBUTION: III. COMPLEX 3D POWER AND PHASE SPECTRA

JEFFREY D. SCARGLE, M. J. WAY^{1,2}, P. R. GAZIS³
 NASA Ames Research Center, Astrobiology and Space Science Division,
 Moffett Field, CA 94035, USA

¹NASA Goddard Institute for Space Studies, 2880 Broadway, New York, NY, 10025, USA

²Department of Astronomy and Space Physics, Uppsala, Sweden

³Thermo Fisher Scientific, San Jose, CA

ABSTRACT

We demonstrate the effectiveness of straightforward but never heretofore used analysis of the complex 3D Fourier transform of galaxy coordinates derived from redshift surveys. Numerical demonstrations of this approach are carried out on a volume-limited sample of the Sloan Digital Sky Survey redshift survey. The direct unbinned transform yields a complex 3D data cube quite similar to that from the Fast Fourier Transform (FFT) of finely binned galaxy positions. In both cases deconvolution of the sampling window function yields estimates of the true transform. Simple power spectrum estimates from these transforms are roughly consistent with those using more sophisticated methods. However we concentrate on the rarely studied Fourier phase spectrum, a simple and general framework for characterizing non-Gaussianity, more easily interpretable than the tangled, incomplete multi-point methods conventionally used. Supported by modern large scale inference methodology, this approach casts a wider net than one-parameter methods based on F_{NL} for example.

Keywords: cosmology: large-scale structure of universe, galaxies: clusters: general

1. INTRODUCTION: PERSPECTIVE AND ASSUMPTIONS

This paper takes the background established by two previous publications on the multiscale structure of the Universe (Way et al. 2011, 2015, Papers I and II, respectively) in a different direction: *direct 3D Fourier analysis of the galaxy positions*. The goal is to maximize simplicity, extraction of information from the data¹ and independence from models or other underlying assumptions. The following list describes the principles underlying this work. These items are largely methodological clarifications and not cosmological assumptions as such. In many cases our approach differs from previous work, references to which are deferred to the next section.

1. **A Limited Cosmological Sample. Testing models against finite-volume data usually involves consideration of *cosmic variance*. To avoid the need to postulate properties of unobserved data that is inherent in this notion, we here adopt a viewpoint nicely described, but not necessarily endorsed, by Peebles (1975a):**

“One can adopt the view that we have only one Universe, that we can see only part of it, and that the analysis ought to be based on that part alone.”

2. **Nearly Noise-free Data.** In our data the uncertainty due to observational errors is essentially negligible (Section 4.3). We thus do not follow the common practice of eliminating irregularities at small scales as noise (or not “topologically persistent”) with smoothing or other informationally destructive practices. Structure detected on all scales yields significant information about the Universe.
3. **Point Distribution, not a Continuous Field.** The galaxies are often discussed in the context of an underlying spatially continuous field – such as an averaged luminous matter density, an underlying dark matter distribution,

Jeffrey.D.Scargle@nasa.gov, Michael.J.Way@nasa.gov, PGazis@sbcglobal.net

¹ Data used herein is the same volume limited SDSS DR7 Main Galaxy Sample (Strauss et al. 2002) used in Papers I and II.

or a random probability density for the presence of a galaxy as a function of position. The current paper addresses the spatial distribution of galaxies without explicit reference to any such continuous fields. We treat galaxies as discrete entities whose spatial distribution carries information about the structure of the Universe and not as tracers of a field. This approach is consistent with equal treatment of all galaxies – i.e. massive galaxies are not given more weight than light ones.

4. **Summary Distributions.** There are two qualitatively different approaches:

- detailed representation and interpretation of local structures
- estimation of a few summary, global distributional parameters

again nicely spelled out in [Peebles \(1975a\)](#). In Papers I and II we opted for the former. Here we derive several Fourier analysis functionals with the goal of estimating important global parameters.

5. **Gaussianity not Assumed a priori.** We approach the search for signs of non-Gaussianity via the Fourier phase spectrum. A complete characterization of Gaussianity is contained in the power spectrum. The information about non-Gaussianity contained in the phase spectrum is organized in a form that is relatively easy to interpret (cf. Section 4.2).
6. **Absolute Clustering.** Much analysis has been carried out on “excess clustering,” treating clumping and spatial correlations as somehow above and beyond an assumed uniform but random background (e.g. [Yu & Peebles 1969](#); [Landy and Szalay 1993](#)). There are in effect two equivalent representations of clustering, absolute and relative to a fiducial background density; statistically these are related by simple transformations. However, our absolute approach avoids some minor numerical problems.² If nothing else we reduce the number of unknown physical processes we need to characterize from two (excess and background) to one.
7. **Explicit Deconvolution of the 3D Selection Function.** Using standard Fourier transform methods we avoid constructs such as Monte Carlo simulations of “un-clustered” points within the selected volume. For example, [Feldman, Kaiser and Peacock \(1994\)](#) state: “Our approach is to take the Fourier transform of the real galaxies minus the transform of a synthetic catalog with the same angular and radial selection function as the real galaxies but otherwise without structure.” Perhaps this approach has meaning in the context of a model based on underlying randomness, but is not a prescription for deconvolving the selection function. Further, in consonance with item 1 and 2, we avoid interpreting such variance as a measure of uncertainty. While various deconvolution methods have been employed for both CMB and galaxy data, we believe the direct 3D Fourier deconvolution described in Section 3.5 is novel.
8. **Use All the Data.** In order to maximize the information gleaned from the analysis, where possible we use all of the data. For example we do not discard galaxies near the edges of the data space in order to simplify the shape of the window function (Section 3.3), but the next item indicates one case where we feel a cut on the data is justified.
9. **Volume-Limited Samples.** Throughout, as in Papers I and II, we use well defined volume-limited samples. This is a minor violation of the previous item, but is for the good reason that it avoids bias corrections necessary for a magnitude-limited sample.
10. **Omitted Effects.** Due to the small radial depth of our relatively shallow volume limited sample ($z \leq 0.12$) and our interest in the simplest analysis, we have neglected many processes known to affect the data, including evolution and nonlinear cosmological terms, peculiar velocities, gravitational lensing, and local and non local GR terms depending on Bardeen potentials and their temporal derivatives (e.g. [Raccanelli et al. 2015](#), especially Fig. 1).

While some of these viewpoints are somewhat nonstandard they are not meant as criticisms of other approaches. Our goal is limited to investigating the simplest possible way to extract spatial frequency information, largely avoiding model-specific assumptions. A case in point: *Cosmic variance*

² Under conventional definitions the correlation function of excess density relative to a constant background integrates to either 0 or $-1/N$. This consequences of this normalization for representation with power laws, which are everywhere positive, is often ignored.

is the uncertainty in parameters estimated from a finite sub-volume of the observable universe. This term refers to the variance among estimates that would be obtained from other sub-volumes.³ The relatively small sample analyzed here does not permit meaningful division of the data into independent sub-samples. Explicit evaluation of cosmic variance in this fashion should be carried out with larger datasets. We thus set aside for future consideration such studies, detailed interpretation, corrections for systematic errors, comparison with models, etc. The simple analyses presented here merely point to some ways in which the complex Fourier spectrum can be cosmologically useful, concentrating more on the phase spectrum and less on the power spectrum. In particular we demonstrate the potential use of kurtosis estimates derived from phase data cubes combined with a relatively new tool for large numbers of parallel statistical tests called *Higher Criticism*.

The organization of the rest of this paper is as follows: Following a brief survey of prior work in Section 2, Section 3 provides explicit details of two different ways to compute Fourier transforms – a direct unbinned approach and a fast Fourier transform of galaxy coordinates in 3D spatial bins – and a simple procedure for deconvolution of the sampling window from the estimated galaxy transform. Section 4 discusses the amplitude (power) spectrum and the phase spectrum as a measure of Gaussianity of the galaxy distribution. An epilog in Section 5 provides a brief summary of the salient results of the paper, followed by two appendices dealing with a few mathematical issues and presenting some of the MatLab code used for the computations.

2. PREVIOUS WORK

A small part of the earlier relevant literature can be found in Limber (1953); Gunn (1965); Yu & Peebles (1969); Peebles (1975a); Peebles and Hauser (1974a,b). The cosmological importance of power spectrum analysis has recently been extensively discussed in Carron, Wolk and Szaudi (2015), especially in the context of galaxy redshift surveys (e.g. Vogeley and Szalay 1996, Sec. 1.1). Fourier phases have been studied in connection with cosmic microwave background data (see e.g. Chiang et al. 2003; Naselsky, Doroshkevich and Verkhodanov 2003, 2004; Chiang, Naselsky and Coles 2004; Naselsky, Chiang, Olesen and Novikov 2005; Chiang and Naselsky 2007; Kovács, Szapudi and Frei 2013; Ferreira and Magueijo 1997). More recently phase analysis is beginning to be applied to galaxy redshift data (Hikage et al. 2005; Matsubara 2007; Wolstenhulme, Bonvin and Obreschkow 2015; Eggemeier et al. 2015).

Some relevant work on non-Gaussianity, much in the context of CMB but with application to large scale – or more appropriately multi-scale – structure includes: Hikage et al. (2006) who provide a general overview, Sefusatti and Komatsu (2007) who discuss the bi-spectrum for high redshift galaxies, Hikage et al. (2008) who discuss the application of Minkowski functionals, Sánchez and Cole (2008) who estimate the power spectrum using Fourier methods based on work by Feldman, Kaiser and Peacock (1994), Martínez-González (2009) and Lentati, Hobson and Alexander (2014) for pulsar timing studies, and also Kovács, Szapudi and Frei (2013). See Coles et al. (2005) for application of Fourier methods to the 2dF galaxy redshift survey, employing the Fourier based method of Percival, Verde and Peacock (2004), a generalization of the minimum variance method of Feldman, Kaiser and Peacock (1994). Coles et al. (2005) derive power spectra and compare them to several empirical (e.g. Tegmark et al. 2004a) and theoretical results. See Kitaura (2012) for derivation of some statistics relevant to non-Gaussianity in galaxy clustering. Other work on non-Gaussianity can be found in (Coles and Chiang 2000; Tegmark et al. 2004b; Rocha et al. 2001).

Efstathiou and Moody (2001) describe a method of recovering the three-dimensional power spectrum from measurements of the angular correlation function applied to the APM galaxy survey – one of the first large surveys using automatic plate measuring methods. See also Querre, Starck and Martínez (2002) for discussion of the galaxy distribution using multiscale methods in general, and 3D implementations of the à trous algorithm, and the ridgelet and beamlet transforms in particular. Percival, Verde & Peacock (2004, hereafter PVP) studied luminosity-dependent galaxy clustering with spherically averaged Fourier analysis. Coles et al. (2005) applied the Fourier based method of PVP to the 2dF galaxy redshift survey. This approach in turn is a generalization of the minimum variance method of Feldman, Kaiser and Peacock (1994). See recent papers (Slepian and Eisenstein 2015a,b,c; Doré et al. 2015) for estimation of three-point correlation functions and their application to problems in dark matter cosmology. Alam, Ata, Bailey et al. (2016) give a recent summary of relevant literature and extensive analysis of data from the SDSS-III Baryon Oscillation Spectroscopic Survey.

³ Rarely, e.g. in discussing the *anthropic principle*, some authors include possible future observations not yet made or not feasible, other parts of the Universe inaccessible to observation, other universes, or a hypothesized random ensemble of characteristics of the early Universe. What is common to all of these is that the data at hand can be regarded as just one realization of a “cosmic” random process.

3. 3D FOURIER TRANSFORMS

The sub-sections below describe the procedures used to compute the complex Fourier transform of the galaxy distribution – first reviewing the data and then outlining direct and binned transforms of the galaxy positions and the corresponding data window, concluding with a Fourier-based procedure to deconvolve the effect of this window function.

3.1. The Data

Below we Fourier analyze the SDSS DR7 data described in Papers I and II, namely the NASA/Ames Research Center SDSS Value Added Galaxy Catalog (AMES-VAGC). Data Release 7 of the SDSS was augmented using the New York University Value Added Catalog (see Appendix A of Paper II for details and references). **As a reminder, redshift ζ , right ascension α and declination δ were converted to rectangular Cartesian coordinates with the formulas**

$$\begin{aligned} x &= \zeta \cos(\delta) \cos(\alpha) \\ y &= \zeta \cos(\delta) \sin(\alpha) \\ z &= \zeta \sin(\delta) \end{aligned} \tag{1}$$

and no cosmological corrections were made in view of the low-redshift nature of the sample. The one small difference is that, since the analysis here does not involve Voronoi tessellation, the omission of the small sample of galaxies near the edges of the data space (Paper I, Section 5.2.2) is unnecessary. This slightly increases the sample size. More significant is the resulting improved definition of the edges of the data space, of importance for the transform of the data window described in Section 3.3.

3.2. Fourier Transform of the Galaxy Distribution

Let's start with the Fourier transform of the data, keeping an eye toward preserving both directional and phase information. For any function $f(\mathbf{x})$, defined over some solid 3D volume V in $\mathbf{x} = (x, y, z)$ -space, the Fourier transform is⁴

$$F_f(\mathbf{k}) = \int_V \mathbf{f}(\mathbf{x}) e^{i\mathbf{k} \cdot \mathbf{x}} d\mathbf{x} \tag{2}$$

where \mathbf{k} is the spatial frequency vector $\mathbf{k} = (\mathbf{k}_x, \mathbf{k}_y, \mathbf{k}_z)$ – chosen so that the linear scale (one full period of the sinusoid) corresponding to k is $2\pi/k$.

Following Yu & Peebles (1969); Peebles (1975a); Peebles and Hauser (1974a,b) we account for the discreteness of the data by taking $f(\mathbf{x})$ in Equation (2) as the sum of point locator functions, i.e. delta functions at the positions \mathbf{x}_n of each of the galaxies:

$$f(\mathbf{x}) = \sum_{n=1}^N \delta(\mathbf{x} - \mathbf{x}_n) \ , \tag{3}$$

where the sum is over the N galaxies included in the volume limited sample. (See also Bardeen et al. (1986) for a similar representation in terms of peaks – local 3D maxima – of density.) The resulting galaxy Fourier transform is simply

$$F(\mathbf{k}) = \int_V \sum_{n=1}^N \delta(\mathbf{x} - \mathbf{x}_n) e^{i\mathbf{k} \cdot \mathbf{x}} d\mathbf{x} = \sum_{n=1}^N e^{i\mathbf{k} \cdot \mathbf{x}_n} \ , \tag{4}$$

or in component notation

$$F(k_x, k_y, k_z) = \sum_{n=1}^N e^{i(k_x x_n + k_y y_n + k_z z_n)} \ . \tag{5}$$

This simple formula is easily evaluated for any galaxy sample, in time of order $N \times N_k^3$, i.e. the product of the number of galaxies and the total number of frequencies (N_k is the number of frequencies in a single coordinate direction). It treats all galaxies as identical points. Through its response to crowding together of galaxies in various regions it is sensitive to the local number density of galaxies, but not directly to mass density. Figure 1 displays the 3D structure of the Fourier power spectrum $|F(k_x, k_y, k_z)|^2$. The higher frequencies roughly speaking form an isotropic but somewhat

⁴ The opposite sign in the argument of the exponential is sometimes used.

irregular shell around the inner core (the black shape at the center of the plot) of low-frequency or large-scale structure. Spectral quantities derived using equations (4) and (5) will be called *direct*, as opposed to *binned* for those from the methods in Section 3.4. Appendix A describes the use of the inverse Fourier transform to check how well Eq. (5) represents the raw data.

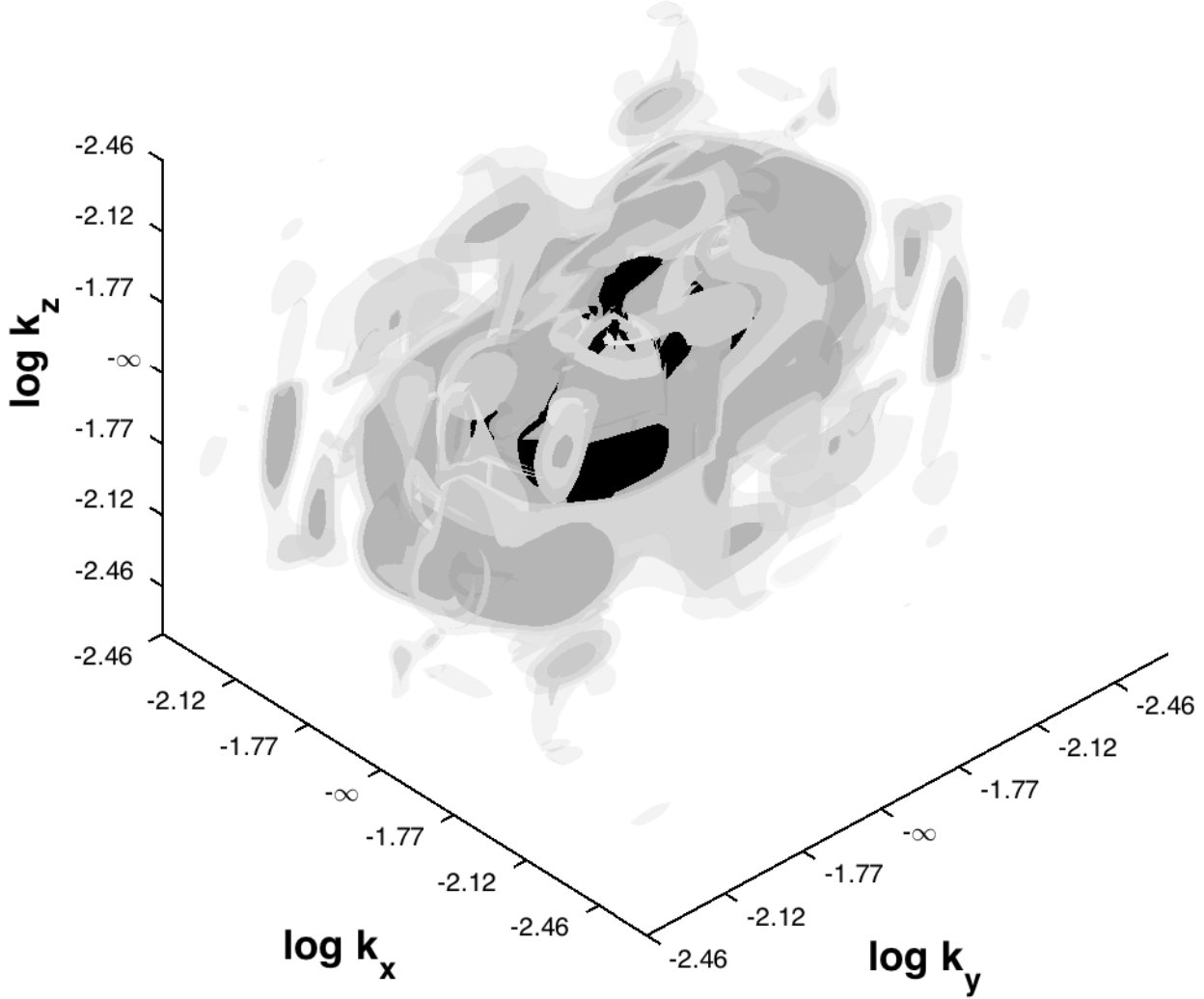


Figure 1. Log-log isosurface plot of the Fourier power spectrum as a function of the x, y and z spatial frequencies. Zero frequencies cannot be plotted logarithmically, as indicated by $-\infty$ at the centers of the axes. Units of k_x , k_y , and k_z are h/Mpc. The powers, in order of decreasing opacity of the iso-surface and expressed as fractions of the zero-frequency power are 0.8357, 0.7612, 0.7450 and 0.7288. These levels were chosen to make this display informative.

3.3. Fourier Transform of the Data Window

The data from a survey of a given volume V can be thought of as the product of the actual 3D spatial distribution of galaxies multiplied by a 3D spatial window, or selection function, given by:

$$S(\mathbf{x}) = \begin{cases} 1 & \text{for } \mathbf{x} \in V \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

This window can be defined by the 2D footprint of the survey on the sky combined with the 1D redshift selection function. Here we use the corresponding volume in terms of rectangular coordinates $\mathbf{x}, \mathbf{y}, \mathbf{z}$. This approach ignores the variation of the redshift selection over the relatively narrow range of our data. Of course any survey has this and additional selections, not considered in this purely geometrical analysis.

By the well known convolution theorem (Bracewell 1999) the Fourier transformation of this product relation yields the fact that the Fourier transform of the survey data [as in Eqs. (2) and (3)] is the transform of the actual distribution convolved with the Fourier window function, defined as the transform of the selection function:

$$F_{\text{window}}(\mathbf{k}) = \int_{-\infty}^{\infty} S(\mathbf{x}) e^{i\mathbf{k} \cdot \mathbf{x}} d\mathbf{x} = \int_V e^{i\mathbf{k} \cdot \mathbf{x}} d\mathbf{x} \quad (7)$$

In order to compute this integral exactly one could discard some of the data and redefine V as a simplified subset of the actual data space, such as a figure with planar boundaries. Here we wish to compute $F_{\text{window}}(\mathbf{k})$ where V is the actual 3D data space of the redshift survey. Note that the linearity of equation (2) means that the Fourier transform can be evaluated as a sum of transforms over the elements of any partition of V ; i.e. for any f

$$F_f(\mathbf{k}) = \sum_n \int_{V_n} f(\mathbf{x}) e^{i\mathbf{k} \cdot \mathbf{x}} d\mathbf{x} \quad (8)$$

where $\{V_n, n = 1, 2, \dots\}$ is a set of disjoint volumes the union of which is the full observation space V . It is convenient here to partition V into a set of rectangular parallelepipeds, or *cuboids*. For then the contribution of each cuboid C^n , i.e. the volume defined by

$$S(\mathbf{x}) = \begin{cases} 1 & x_a \leq x \leq x_b; y_a \leq y \leq y_b; z_a \leq z \leq z_b \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

can be found exactly as a function of its bounding xyz coordinates x_a, x_b etc. The Fourier transform of such a cuboid is just

$$F_{C^n}(\mathbf{k}) = \int_{C^n} e^{i\mathbf{k} \cdot \mathbf{x}} d\mathbf{x} \quad (10)$$

$$= \int_{x_a}^{x_b} e^{ik_x x} dx \int_{y_a}^{y_b} e^{ik_y y} dy \int_{z_a}^{z_b} e^{ik_z z} dz \quad (11)$$

$$= \left(\frac{e^{ik_x x_b} - e^{ik_x x_a}}{ik_x} \right) \left(\frac{e^{ik_y y_b} - e^{ik_y y_a}}{ik_y} \right) \left(\frac{e^{ik_z z_b} - e^{ik_z z_a}}{ik_z} \right) \quad (12)$$

Now let's approximate the data space with a fine partition as follows: Construct a grid of equal squares covering the projection of V onto the x - y plane, with a fine spacing $\Delta = x_b - x_a = y_b - y_a$. To define a cuboid it remains to specify z_a and z_b . We take these as the z -coordinates at which a vertical line through the center of the square and parallel to the z -axis intersects the convex hull of the full set of galaxy positions. It is easy to see that each such line intersects the convex hull in either 2 or 0 facets; in the latter case the cuboid is entirely outside the data set and is ignored.

Figure 2 shows relatively crude partitions of the actual data space with the long axes of the cuboids in two different directions. If the transverse dimensions of the cuboids are sufficiently small the partition approaches an exact coverage of the overall convex hull and the resulting window Fourier transform is independent of the cuboid orientation. For the grid size equal to .0001 redshift units (0.416 Mpc) the sum of the volumes of the cuboids matches the exact volume of the convex hull to one part in 10^8 , not surprising since this computation is equivalent to the elementary integral calculus procedure for computing the volume of the convex hull. Putting all this together, the result will be used to correct the galaxy Fourier transform for the effects of the data window – cf. Section 3.5. Appendix A describes a way to check how well the transform approximates the actual selection function, and Appendix B gives some MatLab code relevant to computing the transform of the selection function.

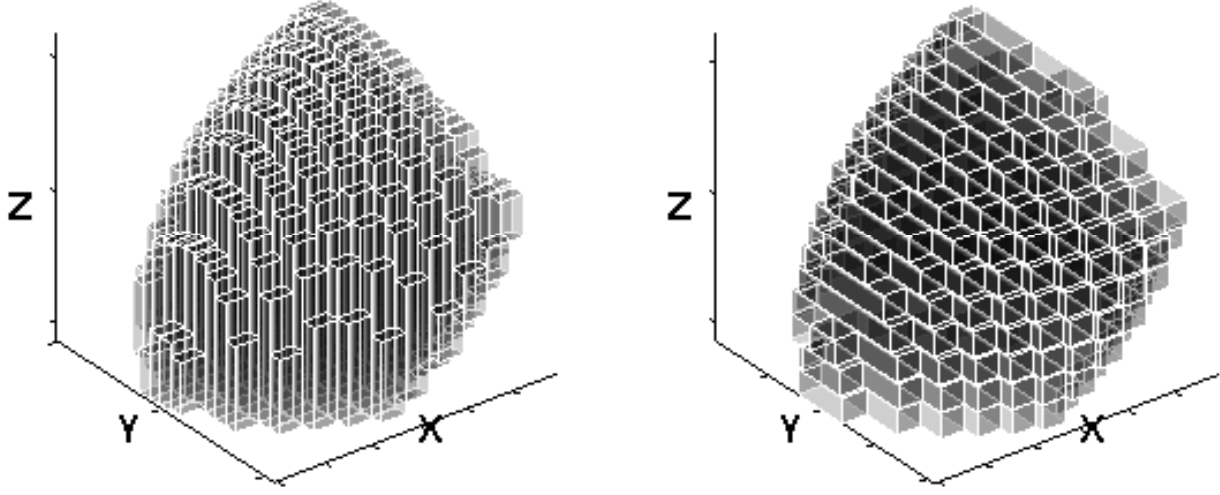


Figure 2. Two coarse partitions of the convex hull of the SDSS data using cuboids with transverse size .01 redshift units. The long-axes of the cuboids are, left to right, parallel to the z-axis, and y-axis, respectively.

3.4. Fast Fourier Transform of the Binned Galaxy Distribution

Another way to estimate the Fourier transform is to construct a 3D histogram of the galaxy positions in a spatial grid of 3D bins, or *voxels*. To use the fast Fourier transform each voxel must be a cube with the same small size $\Delta x = \Delta y = \Delta z$. This procedure discards some information, due to rounding of galaxy coordinates and placing some close pairs in the same voxel. Table 1 summarizes statistics for three grid sizes. Columns 4, 5 and 6, giving the maximum number of galaxies in any one voxel and the fractions with 1 galaxy and 2 or more galaxies, are useful in assessing the information loss in this binning. Ideally the fraction with more than one galaxy would be zero, leaving coordinate truncation as the only error. Computation with 512 bins in each coordinate – case (c), with 135,005,697 voxels – seems to be the largest feasible with current personal computers. As the accuracy of the computation increases more and more voxels are empty, since the number not empty cannot exceed the number of galaxies. The last column gives the fraction of voxels empty because they are outside the survey volume; these values are large due to the shape of the survey and because we zero-padded it for good frequency resolution.

Table 1. Statistics for the Binned Fourier Transform

Case	$N_{\text{bins}}^{\text{a}}$	$N_{\text{pix}}^{\text{b}}$	Δ^{c}	Max n ^d	Fraction n=0 ^e	Fraction n=1 ^f	Fraction n > 1 ^g	Fraction Outside ^h
(a)	128	2.1 M	6.8	32	0.9693	.016519	.014221	0.8923
(b)	256	16.8 M	3.4	11	0.9938	.004882	.001300	0.9347
(c)	512	134.2 M	1.7	5	0.9991	.000869	.000062	0.9347

^aNumber of bins per dimension.

^bNumber of voxels (millions).

^cLinear dimension of voxels (Mpc).

^dMaximum number of galaxies in a voxel.

^eFraction of empty voxels.

^fFraction of voxels containing one galaxy.

^gFraction of voxels containing more than one galaxy.

^hFraction of voxels outside the convex hull of the data.

The Fourier transform of the window is simply that of this bin array with unity inside the sample volume and zero outside, cf. equation (7). Actually instead of the convex hull of the filled bins, for each dimension we assigned a unit value to each bin between the minimum and the maximum indices of bins containing galaxies in all of the corresponding x-columns, y-columns and z-columns. In practice this is essentially the same as the convex hull. The inverse transform of the Fourier transform computed this way is guaranteed to exactly reproduce the input counts-in-voxels, so there is no point in numerically demonstrating the accuracy of this representation as in Appendix A for the direct transform.

3.5. Deconvolution of Windows

We approach correcting for the selection function (or window) in a straightforward way. Any pair of functions of a 3D coordinate vector \mathbf{x} related multiplicatively,

$$q_{obs}(\mathbf{x}) = q_{true}(\mathbf{x})q_{window}(\mathbf{x}) , \quad (13)$$

have spatial Fourier transforms related by

$$Q_{obs}(\mathbf{k}) = Q_{true}(\mathbf{k}) * Q_{window}(\mathbf{k}) \quad (14)$$

where $Q_{obs}(\mathbf{k})$ is the Fourier transform of $q_{obs}(\mathbf{x})$, etc., and $*$ means 3D convolution on the vector \mathbf{k} . There are many *deconvolution* techniques for solving such equations for $q_{true}(\mathbf{k})$, thus correcting for the window function. Here the obvious Fourier transform solution suffices:

$$q_{true}(\mathbf{k}) = \mathbf{F}^{-1} \frac{\mathbf{F}[Q_{obs}(\mathbf{k})]}{\mathbf{F}[Q_{window}(\mathbf{k})]} \quad (15)$$

where F and F^{-1} are the 3D direct and inverse Fourier transforms, respectively. In all numerical results presented here the MatLab (©MathWorks) multidimensional functions `fftn` and `ifftn` were used for both the direct and binned cases. This deconvolution method is sometimes avoided because of worries about noise amplification and/or issues when the denominator in eq. (15) is zero (or small in absolute value), but here these issues have not caused any serious problems.

4. CHARACTERIZING THE SPATIAL DISTRIBUTION OF GALAXIES

We are now ready to use the above Fourier transform methods to globally characterize the scale-dependence and position-dependence (via the power spectrum and phase spectrum, respectively), and Gaussianity of the galaxy number-density distribution. This will be accomplished by quantifying these properties as they are manifested in the volume-limited data sample at hand. Following the ideas of Section 1, to analyze the variance of such parameters (beyond that due to the small positional measurement uncertainties) would require extrapolation to hypothetical data outside the sample at hand.⁵

In the following it is useful to compare results from the binned and unbinned Fourier transforms. Neither one is better in all aspects than the other. Of course they both have limited spatial frequency resolution, but their different data representations implement distinct approximations. The binned approach suffers from the information loss associated with quantization of the galaxy coordinates.

4.1. Fourier Power Spectrum

Figure 3 shows the deconvolved power spectra for both methods: direct as in Sections 3.2 and 3.3 and binned in Section 3.4. The powers projected in three orthogonal directions, $|F(k_x, 0, 0)|^2$, $|F(0, k_y, 0)|^2$, and $|F(0, 0, k_z)|^2$, are distinguished by lines of different widths. These three power spectra share the same zero-frequency value, namely $|F(0, 0, 0)|^2 = N^2$, and we have normalized the plotted curves to unity at zero spatial frequency – which of course is off-scale on these log-log plots. Comparison of the power spectra in different directions provides a simple measure of isotropy. The spectra at lower spatial frequencies approximate the power-law dependence characteristic of red noise (Aschwanden 2011). The straight (dashed) lines in this figure are least-squares fits to the mean of the three power spectrum curves in the interval below the cutoff at $\log k = -1.2$ mentioned in the caption; the log-log slopes indicated there are not far from the common red noise value of ≈ -2 . The scatter reflecting high variability at small scales motivates the vertical shifts in the higher frequency part of these plots, at the same cutoff used for the power-law fits.

⁵ These are then *spatially local quantities* referring to the subset of the observable Universe contained in the data sample. The quandary is much like one that arises in time series: it is impossible to establish with certainty – from a single finite realization of data – whether or not a time series process is rigorously stationary. One can nevertheless construct measures of *local stationarity*.

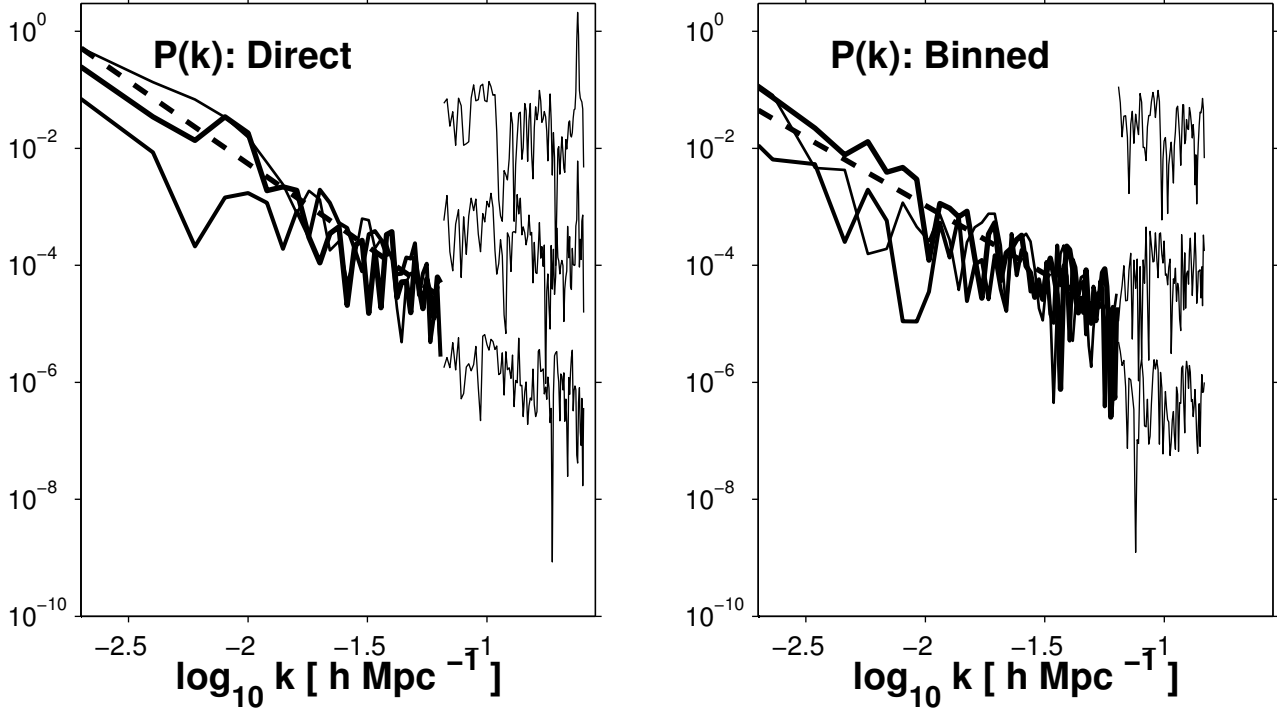


Figure 3. Power Spectra from deconvolved direct (left) and binned (right) Fourier Transforms: x, y and z powers in solid lines of increasing width. Above $\log k$ of -1.2 (spatial frequencies > 0.063) the powers are multiplied by 3000, 10, and 0.1, respectively, for clarity. The dashed straight lines are power law fits to the low frequency data (averaged over the 3 directions) with slopes -2.8 and -2.3 respectively.

Figure 4 plots our power spectra against those from some other authors. In interpreting the figure and assessing this comparison the reader should bear in mind both the simplicity of our method – using the unadorned Fourier basis and avoiding the variety of known weighting schemes, corrections, and assumptions – and the differences in the data used. This figure compares the average of our three x, y and z projected spectra in Figure 3 with results from detailed analysis of very similar data by Tegmark et al. (2004b) and of a much larger sample by Percival, Nichol, Eisenstein et al. (2007).

Using a flux limited sample, instead of our more easily interpreted volume limited sample, the first authors address the selection function, redshift space distortions, bias effects and other systematic errors, using a Pseudo-Karhunen-Loeve expansion (Tegmark et al. 1998). Figure 4 includes the data from the first two columns of their Table 2 in the form of open circles, without showing their rather large horizontal and vertical error bars. They refer to this as the real-space galaxy-galaxy power spectrum P_{gg} in units of $(h^{-1} \text{Mpc})^3$, and “recommend using column 2 for basic analysis.” Like ours this estimate treats the galaxies as equal points and accordingly is not corrected for bias, justified because bias appears to be largely luminosity and scale independent (their Figures 28 and 29, renormalizing to the linear Λ CDM model). It is noteworthy that their power spectra with and without correction for the Fingers of God (FOG) – their columns 2 and 3, respectively – would be indistinguishable had we plotted both. Even though the effect of FOGs seems insignificant here redshift space distortions should be addressed in any serious scientific applications. Analysis of a much larger redshift survey, extending to much larger redshifts than our study, by Alam, Ata, Bailey et al. (2016), includes significant redshift-space distortion corrections. The results of a similarly detailed analysis by Percival, Nichol, Eisenstein et al. (2007) of a sample including both SDSS main galaxies and luminous red galaxies (LRGs) out to much larger redshifts ($z \sim 0.5$) than our sample, are plotted as plus-signs.

First compare the curves for the direct and binned transforms (solid and dot-dash lines). While the values at some frequencies, especially the lower ones, differ by nearly an order of magnitude, the rough similarity of the slopes and values at higher frequencies demonstrates that these two methods are **crudely** consistent with each other. As well, the similarity of some of the finer detail in the two representations support the notion that the effective spatial resolution is relatively good (probably better than that corresponding to Tegmark et al.’s horizontal error bars, not shown here). The differences between our spectra and the others, especially in the form of a vertical offset above about $k \approx .05$, are

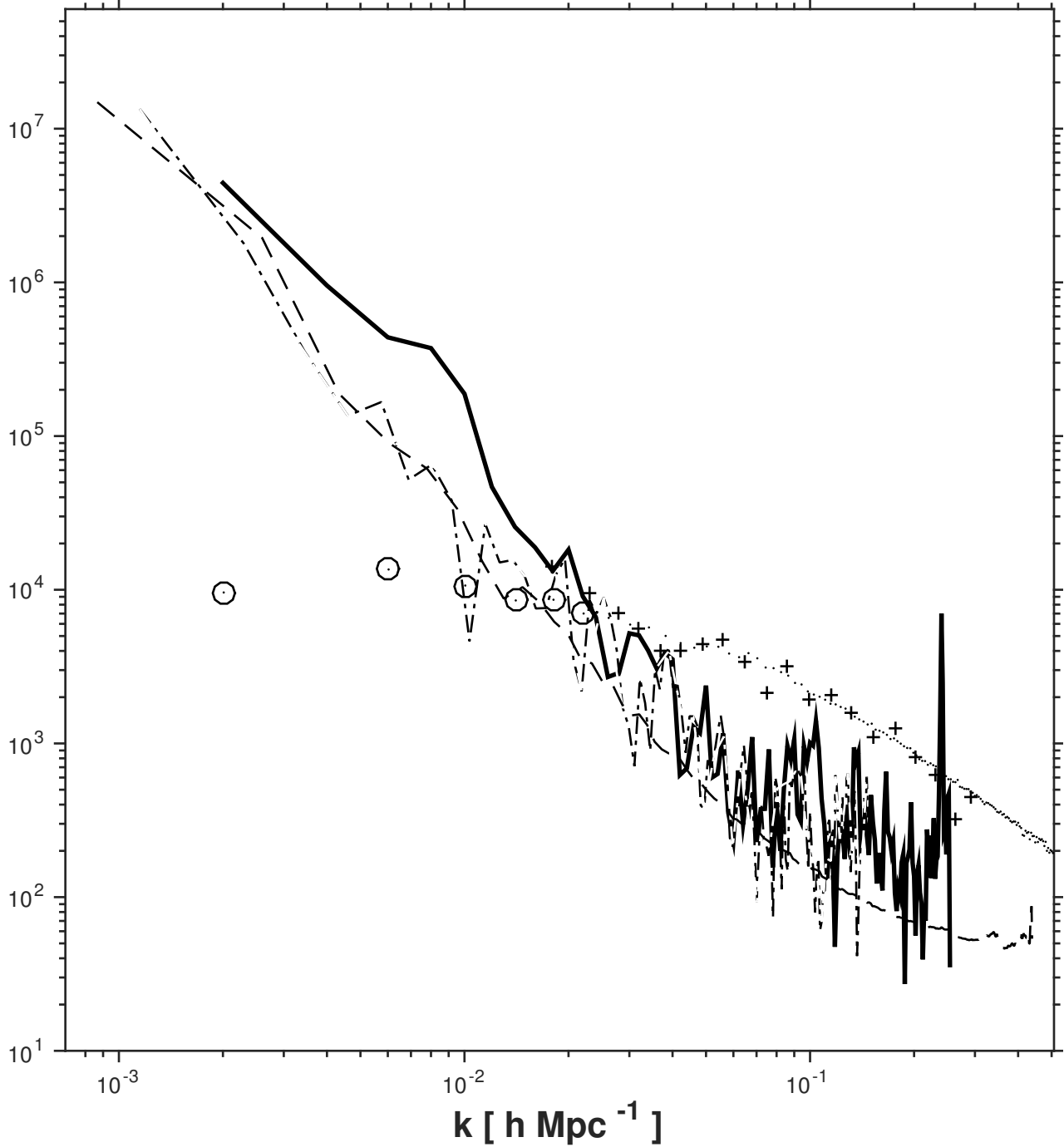


Figure 4. Power Spectrum Comparison. Solid line: average direct Fourier transform (eq. 5). Dots-dashes: average binned FFT. Dashes: is the average of the powers at all of the spatial frequencies falling in a given 1D spherical volume in \mathbf{k} space. These spectra are all corrected for the selection window (cf. Section 3.5) and the curves are plotted with small scaling factors imposing agreement at zero frequency. The spatial frequencies and powers from columns 1 and 2 of Tegmark et al. (2004b) are plotted as plus signs (+), and those of Percival, Nichol, Eisenstein et al. (2007) as small dots (the first 6 frequencies emphasized with circumscribed circles). These spectra are projected to 1D in essentially the same fashion as for our solid line.

not surprising in view of the differences in the data and methods used.

Nominally the plotted points in our power spectra are independent of each other. Essentially no significant measurement errors propagate into this plot at any spatial frequency. The only large discrepancy in the plot is between Percival et al.'s and our powers at the longest scales, understandable in terms of the difference of the data samples and systematic effects at large scales. In the power spectra in Tegmark et al. (2004b) and ours, not surprisingly there is no evidence for baryon acoustic oscillation features. These important features do begin to appear at around $k = 0.7 \text{ Mpc}^{-1}$

with the larger sample and the inclusion of the SDSS luminous red galaxies in (Percival, Nichol, Eisenstein et al. 2007).

4.2. Gaussianity

In modern cosmology it is often posited that the initial conditions of the Universe consisted of a random density field described as a *Gaussian random field*. **It is not completely clear how the character of such an initial distribution may have evolved gravitationally, or how matter-to-galaxy biasing, integrated Sachs-Wolfe (ISW) effects, and gravitational lensing may complicate conclusions based directly on the galaxy distribution (Coles 2000).** Hence the interpretation of detected non-Gaussianity (NG) in the distribution of low redshift galaxies would not be straightforward. We find no NG signatures here, but if significant detections were to be made, e.g. in future large redshift surveys, the resulting parameters would be useful as additional constraints on precise cosmological evolution models. Hence we now describe some aspects of direct analysis of the Fourier phase spectrum.

Although Gaussian processes are well understood mathematically, the elusive nature of *non-Gaussian* (NG) processes has complicated and discouraged exploration of such searches. The infinite number of ways a process can depart from Gaussianity leads to a plethora of potential NG tests, only a handful of which have been pursued. Here we describe a relatively straightforward class of tests through analysis of the complex Fourier data cube. The idea centers around metrics of how identically and independently (IID) the Fourier phases at different spatial frequencies are distributed.

Much previous work centers on parametric tests, valid only in the context of hypothetical physical or mathematical models and thus far short of general characterization of NG. Analyses using higher-order spectra and correlation functions, cumulants, or function bases such as Karhunen-Loeve expansions (Vogeley and Szalay 1996; Tegmark et al. 2004b) or harmonic oscillator eigenfunction expansions (Rocha et al 2001) are closer to the spirit of non-parametric analysis with its greater generality and flexibility. On the other hand these methods are simply ad hoc ways to project an infinite dimensional function space into lower dimensions for modeling convenience, and may suffer fundamental problems (Ferreira and Magueijo 1997; Carron 2011; Carron and Szapudi 2015; Wolk, Carron and Szapudi 2015, e.g.). By contrast the approaches of Rocha et al (2001) and Contaldi et al. (2000) employ Bayesian frameworks that alleviate some of this ad hoc character. But the conclusions are still dependent on the correctness of a hypothetical model (e.g. the quantum mechanical harmonic oscillator in Rocha et al (2001)). More recently Kovács, Carron and Szapudi (2013) define *generalized phases* and apply this concept to characterize the coherence between WMAP and Planck CMB maps.

Various authors have made suggestions, mostly in the CMB context, for the two aspects of this problem, namely identification of: (a) phase subsets that are computationally practical but do not discard too much information, and (b) non-Gaussianity metrics for these sets (Chiang et al. 2003; Chiang, Naselsky and Coles 2004; Naselsky, Chiang, Olesen and Novikov 2005; Chiang and Naselsky 2007). For one example Chiang, Naselsky and Coles (2004) discuss a number of general problems and propose an innovative procedure using return maps. This can be thought of as a way to quantitatively characterize joint distributions (cf. Scargle 1990, e.g.). In another example Chiang and Naselsky (2007) propose ring-like sets in spatial frequency space. And more recently several authors have proposed phase analysis based on 3-point correlation functions of the Fourier transform of a whitened version of the density field (Wolstenhulme, Bonvin and Obreschkow 2015; Eggemeier et al. 2015).

4.2.1. The Fourier Phase Spectrum

We utilize the Fourier phase spectrum as a convenient setting for quantitative characterizaion of the Gaussianity of the galaxy distribution. Several considerations motivate this approach:

- (1) The phase spectrum captures much of the information on non-Gaussianity present in the data.
- (2) Phases of Gaussian data are identically and independently distributed (IID; see e.g. Naselsky, Chiang, Olesen and Novikov 2005). Measures of dependency in the phase distribution are consequently measures of non-Gaussianity.
- (3) As mentioned above the commonly used bi-spectrum and higher order spectra and correlation functions are known to suffer from incompleteness, redundancy, impractical complexity, and severe information loss.
- (4) In most practical, largely non-astronomical, situations the Fourier phase information in 2D images is much more important than the amplitude information (Oppenheim and Lim 1981; Coles 2000; Chiang and Coles 2000). See also (Mannell 1990) for a related discussion of phase in speech intelligibility.

Coles (2000) gives a clear discussion of the background for these points, to which we add only a few remarks. Of course the basic notion is that the Fourier Transform appraises structure as a function of scale. The discrete estimate is a finite sampling of a potentially infinite number of degrees of freedom. But the Nyquist-Shannon sampling theorem guarantees that it captures all the information contained in the data, limited only by the data resolution. Since the inverse Fourier transform exactly recovers the raw data, it is clear that the (frequency dependent) amplitudes and phases contain complementary information, together yielding a complete description of the data. The Fourier *power spectrum* completely characterizes the Gaussian properties of the data; while non-Gaussianity information can appear in both amplitude and phase spectra, in many situations the latter dominates.

Driven by these comments our basic approach is to measure NG through the departure of the phases from being IID. A simple and straightforward approach is to examine the distribution of the set of all phases for nonuniformity. If the Fourier modes were independent, as in true Gaussianity, averages of phases in any set of spatial frequency bins would be the same, but a small amount of non-Gaussianity could show up in this distribution. Figure 5 shows simple histograms of all 16 million-plus phases for the four cases, with $256+1$ spatial frequencies in all 3 dimensions. There is no evidence for any departure from uniformity. If these plots were scaled to include the zero of the ordinate the fluctuations would be invisible. It is notable that even the distributions for the phases not corrected for the data window (the first and third panels) show no such evidence. **This somewhat surprising result, which carries over to subsequent results, no doubt reflects that the data window truncates the Fourier components but does not change their phases.**

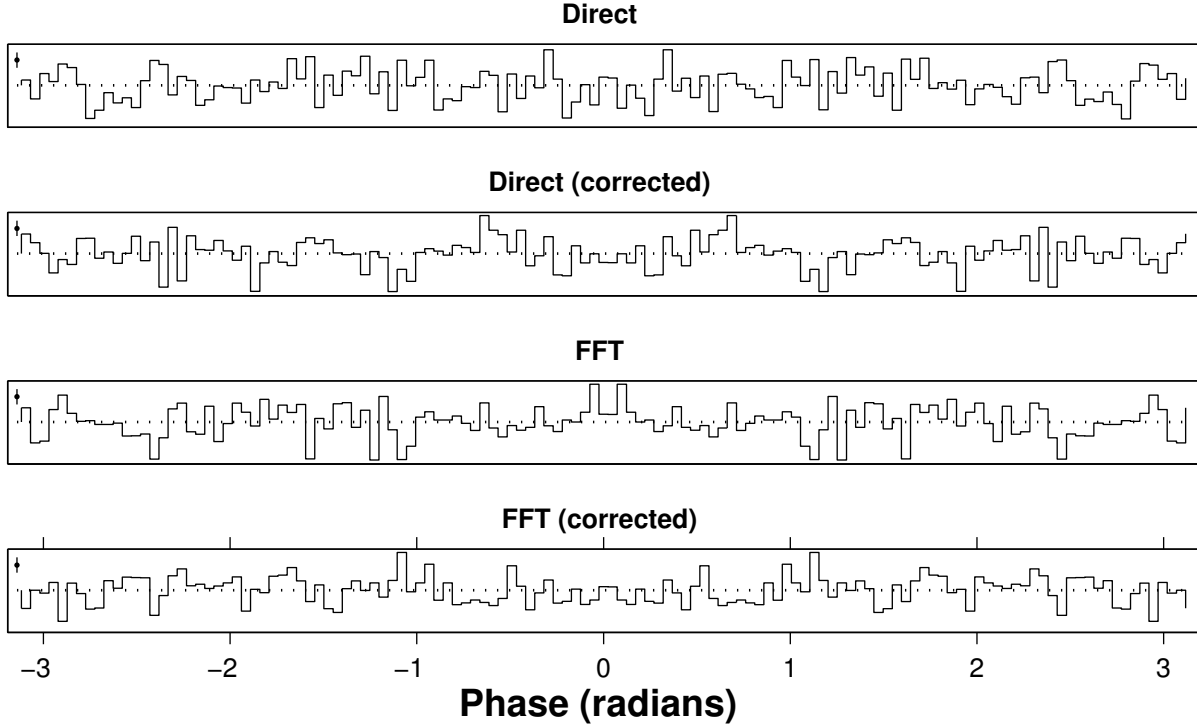


Figure 5. Distributions of phases for the four cases (Direct as in eq. (5), and simple FFT of binned data as in Section 3.4, both with and without correction for the window function, as labeled). Horizontal axis is phase in radians. The vertical axis is the histogram population (128 bins) with a horizontal dotted line at the expected rate of $256^3/2\pi = 2,670,176.86$. Poisson count error bars for a typical bin are shown in the upper left corner of each plot.

The distributions in Figure 5 take no account of frequency relationships, and are thus very insensitive to most of the sorts of departures from IID one might reasonably expect. Ideally one should test not just this single 1D distribution but *joint distributions* of subsets of the phases. The definition of an IID random process X_n is that its joint distributions of all finite orders factor into the product of the corresponding individual distributions:

$$P_m(X_{n_1}, X_{n_2}, X_{n_3}, \dots, X_{n_m}) = \prod_{k=1}^m P_1(X_{n_k}) \quad (16)$$

for all finite m and all choices for $n_1, n_2, n_3, \dots, n_m$. P_m is the m -th order joint distribution function. This condition is much stronger than that the X_n are uncorrelated (e.g. Scargle 1981) – IID implies uncorrelated, but not vice versa.

Hence testing for identically and independently distributed phases is much preferred to tests for merely uncorrelated phases. However an IID metric rigorously and fully implementing this definition is not practical.

We are aided in defining a feasible approximation scheme by the model-independent and non-parametric way the phase spectrum neatly lays out the relevant information in a data cube. We need to identify sets of frequencies related to each other in some germane way. For example, phases at nearby frequencies might show dependencies when those at well-separated frequencies might not. Even this approach is only part of the picture, for *phase differences* also convey relevant information (Coles 2000). One could imagine any number of other constructs potentially useful as metrics of IID phases. The formalism adopted here for NG phase analysis consists of five steps:

- (1) Compute the complex 3D Fourier transform $A(k)e^{i\phi(k)}$.
- (2) From this transform compute the 3D data cube of phases ϕ as a function of (k_x, k_y, k_z) .
- (3) Specify a collection of subsets of this data cube to be tested.
- (4) Select an IID metric; compute it for each of the subsets in this collection.
- (5) Assess the statistical significance of the results of this collection of tests.

Generally Item 4 is expressed as a test statistic (TS), used to test the null hypothesis of IID phases.

The first two steps are straightforward from the discussion in Section (3). The choice in step (3) is most conveniently specified through a condition on the associated spatial frequencies. One example would be a test based on the distribution of differences between phases at adjacent frequencies (Coles 2000). Consider the following example of the many possible ways to take advantage of the organized way frequencies are arranged in a 3D phase-data cube. Let N_k be the size of the Fourier transform in each of its 3 dimensions. For fixed values of each of the N_k^2 pairs (k_y, k_z) compute test statistic T_x for the array consisting of the corresponding N_k phase values as a function of k_x .

4.2.2. Skewness and Kurtosis

The next step is to select a test statistic TS generate a collection of estimates of it, in the form

$$TS_x(k_y, k_z) = T_x \phi(k_x, k_y, k_z) \quad k_y = 1, 2, \dots, N_k; k_z = 1, 2, \dots, N_k \quad (17)$$

(TS for test statistic; the notation T_x indicates which variable T operates on). The Planck Collaboration studied both *skewness* and *kurtosis* Ade et al. (2014) in the role of a test statistic for NG. These are measures of asymmetry of a distribution and of the relative importance of the center versus tails. Jin et al. (2005), based on detailed study of various methods of CMB non-Gaussianity detection, concluded that analysis of the kurtosis of wavelet coefficients is best. The rest of this section describes our approach to using both of these statistics to probe Gaussianity of the 3D phase spectrum.

Table 2. Statistics of Skewness and Kurtosis for Normal and Uniform Variables

Name	Definition	Mean (Normal)	σ (Normal)	Mean (Uniform)	σ (Uniform)
Skewness	$\gamma(\phi) = \sqrt{\frac{n}{6}} \frac{\frac{1}{N} \sum_i (\phi_i^3)}{(\frac{1}{N} \sum_i (\phi_i^2))^{3/2}}$	0	1	0	.5879
Excess Kurtosis	$K(\phi) = \sqrt{\frac{n}{24}} \left[\frac{\frac{1}{N} \sum_i (\phi_i^4)}{(\frac{1}{N} \sum_i (\phi_i^2))^2} - 3 \right]$	0	1	$-\frac{6}{5}$.2421

Table 2 gives definitions following the normalization conventions in Jin et al. (2005) and assuming the mean value of ϕ has been removed. The form here is called *excess kurtosis* because subtracting 3 yields an expected value of zero for the kurtosis of a normally distributed variate. We use this form of the statistic throughout, even though its expected value is not zero for the case relevant here – uniformly distributed variates.

Unwrapping the phase values along the projected axis, e.g. the k_x axis in eq. (17), might enhance the NG signal-to-noise ratio; on the other hand the non-random structure imposed by unfolding significantly affects the data distribution.

Experiments on wrapped and unwrapped phase arrays (not reported here) did not yield any interesting differences. The relative effectiveness of these two approaches presumably depends on the nature of the suspected NG process. This concept should be kept in mind, especially for future studies in the context of specific NG models, but here we ignore it.

Consider now a sequence of examples, starting with idealized synthetic data, adding various features one at a time leading to more realistic situations, and ending up with the actual data. It is important to carefully assess the effects of features of the distribution of the points in synthetic cases, and the boundaries of the data space in all cases. Within all of the following figures, each of the panels is a different view of the same data, synthetic or actual. When the data fall within the irregularly shaped volume we have used different spatial frequency arrays in the three directions. That is, the frequencies are integer multiples of a fundamental frequency defined by

$$k_0(n) = \frac{2\pi}{L(n)} \quad (18)$$

where $L(n)$ is the range of the data in coordinate direction n . This choice seems to give slightly better deconvolutions. In contrast, for convenience the power spectra presented above in Section 3 refer to the same frequency array in all directions, namely corresponding to the largest of L_x , L_y , and L_z .

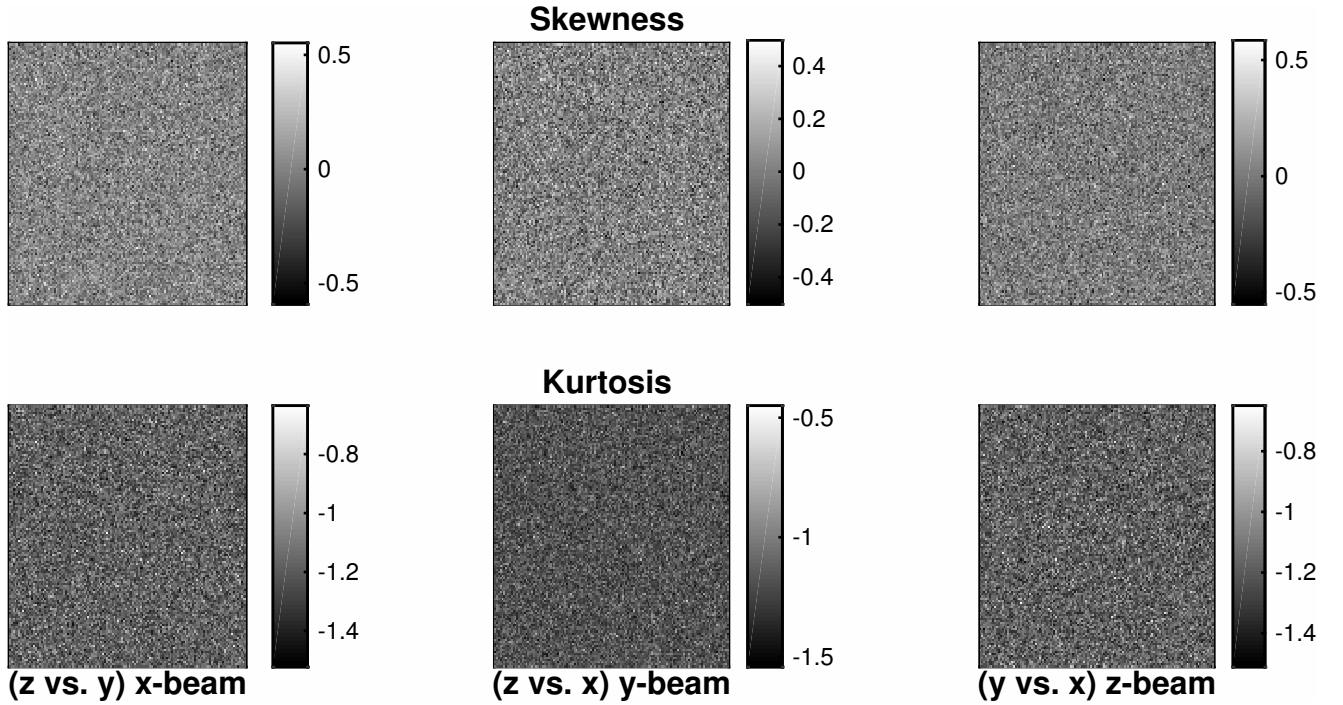


Figure 6. Maps of NG statistics for random phases. All images are from the same 3D $128 \times 128 \times 128$ cube of data consisting of numbers uniformly distributed on $(0, 2\pi)$. Columns from left to right: beams in the x, y and z directions. Top row: skewness; Bottom row: kurtosis. Coordinates are indices in the synthetic random arrays, not functions of spatial frequency as such, so axis labels are suppressed. Here and in subsequent figures the grayscale bars to the right of each panel show how the values spread around the nominal 0 for skewness and -1.2 for kurtosis.

Figure 6 contains maps generated from 3D random phase cubes. Such images are used throughout this section to search for possible non-random patterns in the behavior of NG statistics, computed along beams parallel to the coordinate axes, as a function of coordinates perpendicular to the beams. The first column displays in gray scale the statistic computed along the x-direction, as a function of y and z; similarly for the other two columns, as labeled. These unsmoothed plots retain the discrete nature of the data⁶ to allow better appreciation of the randomness of the distribution. The data cube generating this figure consists of variables generated from an IID random number

⁶ Specifically we used the MatLab `flat` mode for shading plots, not the interpolation mode `interp`.

generator, not from a Fourier transform, and thus represents the most extreme form of the null hypothesis of *random phases*. There is no apparent structure in any of these panels – skewness or kurtosis.

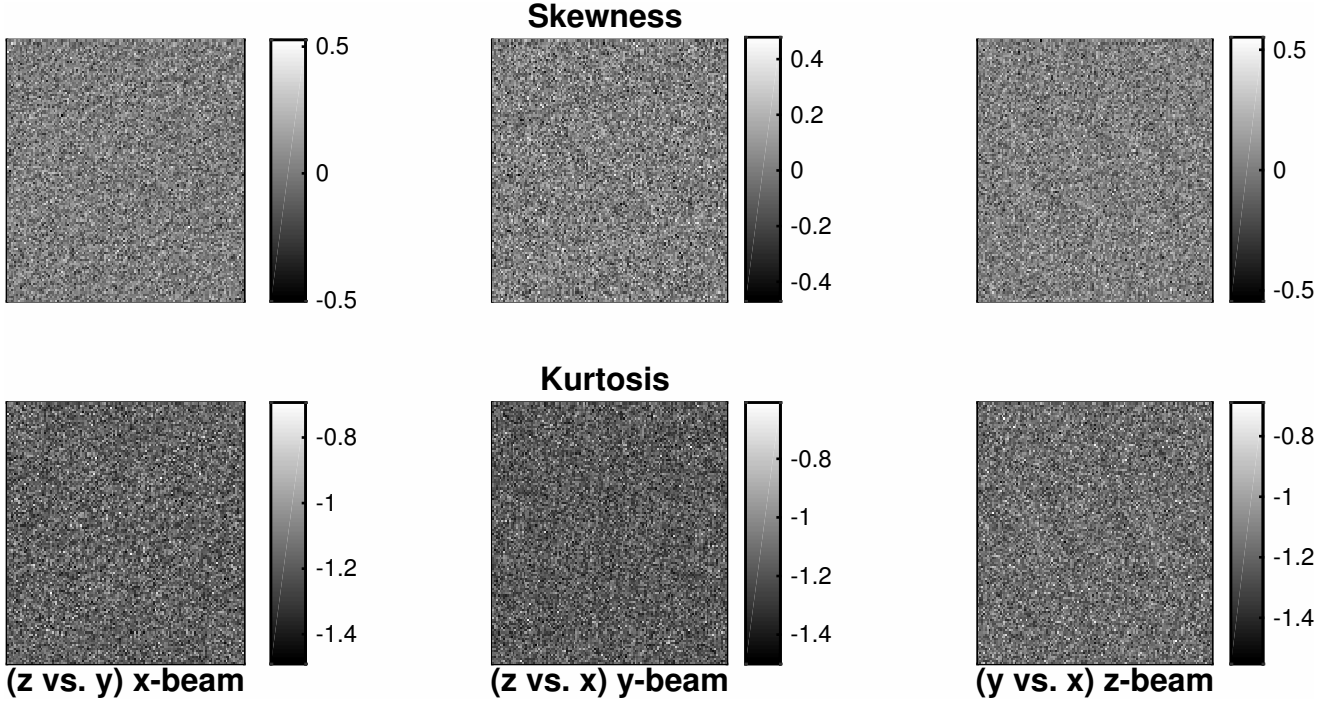


Figure 7. NG statistics maps for phases of the direct Fourier transform of a set of 100,000 xyz points randomly and uniformly distributed within a cubic 3D volume, using equation (5). The identities of the panels are as in Figure 6. While the coordinates are now spatial frequencies, the units are fixed by the arbitrary size of the cube, and therefore are also arbitrary. The 128 frequencies shown here cover the range $-f_0$ to f_0 , where $f_0 = 2\pi/L$ is the *fundamental frequency* and L is the cube size. Zero frequency is the point at the very center of the plot. These axis scales apply as well to the subsequent figures.

Figure 7 shows the same NG maps as in Figure 6 but now the phases are derived from a direct Fourier transform of random points distributed uniformly within an xyz cube. This configuration is chosen to diagnose possible modification of the phase distribution inherent in the transform procedure, but with a benign window due to the simplicity of its boundaries. These figures show a few very weak wave-like patterns, perhaps more prominent for kurtosis than for skewness. Studies with increasing numbers of points show that these patterns become weaker as the number of random points increases, and are therefore deemed to be the result of random fluctuations that decrease with increasing N . Deconvolution is assumed to be unimportant due to the simplicity of the boundaries of the data space.

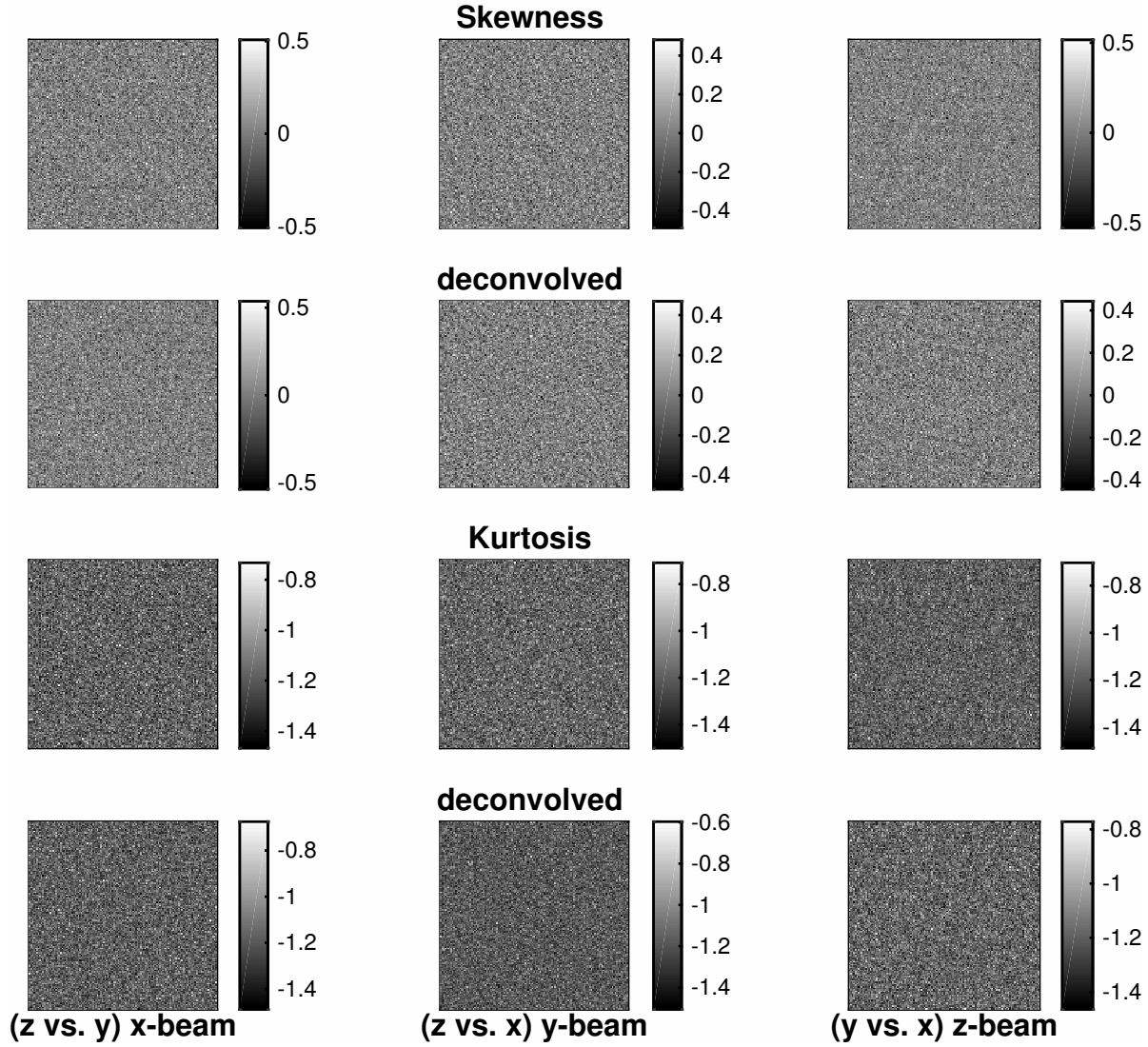


Figure 8. Skewness and kurtosis statistics maps of phases from the Fourier transform of 139,798 xyz points (the same as the number of galaxies in our SDSS data set) randomly distributed within the convex hull of the actual data. Columns are the 3 projections as in previous figures. Rows 1 and 3 are skewness and kurtosis, respectively, computed from the phases of the Fourier transform; rows 2 and 4 are for the same statistics computed from phases of the deconvolved Fourier transform.

Figure (8) presents phase analysis for xyz-data that are still synthetic random points but now distributed uniformly within the convex hull of the actual data. The idea is to diagnose possible structure in these maps induced by the irregular boundaries of the data space. The lack of structure here indicates that such distortion is minimal.

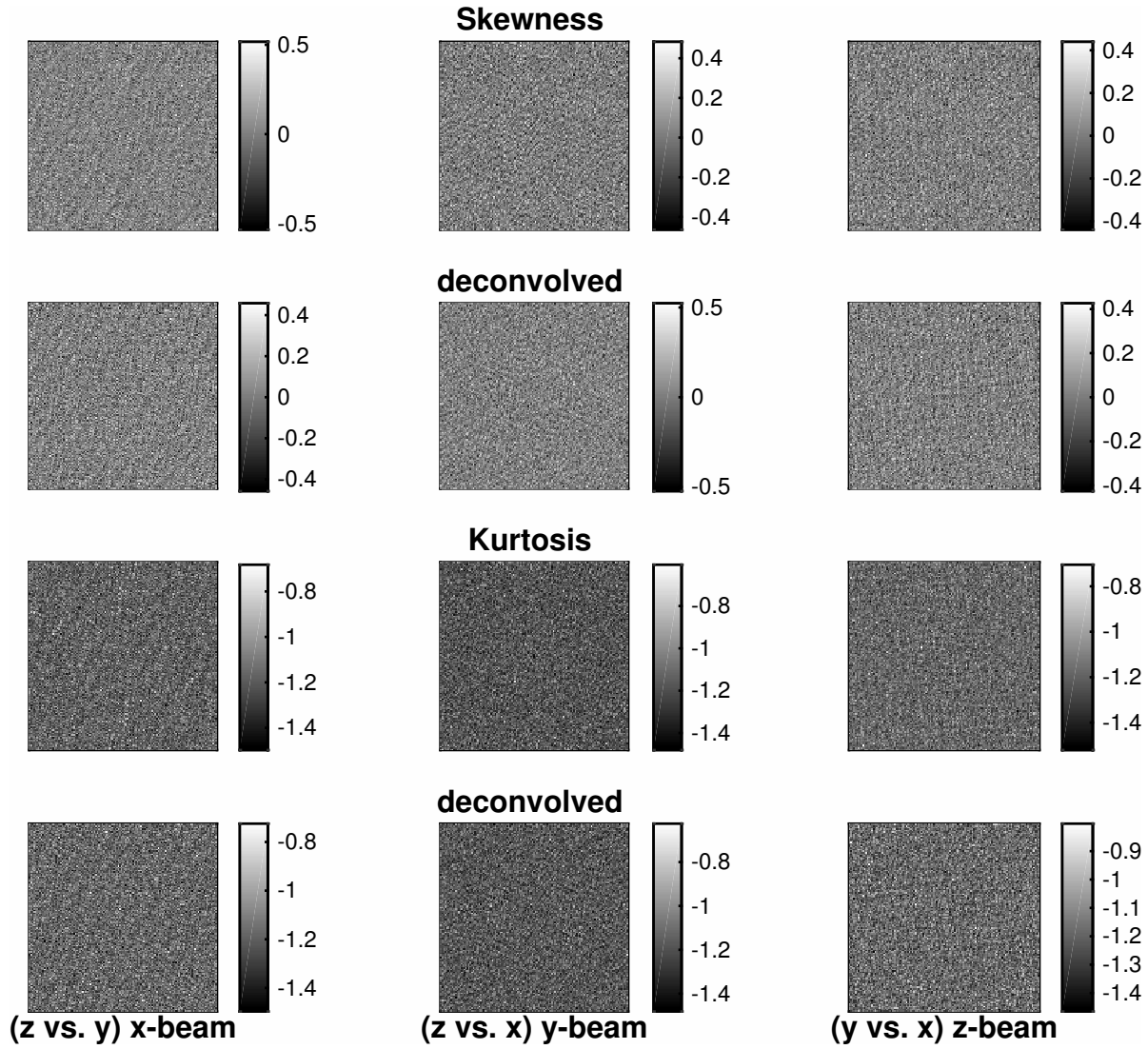


Figure 9. Skewness and kurtosis statistics maps of phases from the Fourier transform of the actual 139,798 xyz points, displayed in the same way as in Figure 8.

Figure (9) presents analysis of the skewness and kurtosis of the phase data cube for the Fourier transform of the actual xyz-data. Comparison of this and Figure (8) indicates that no significant NG effects are present in the galaxy data.

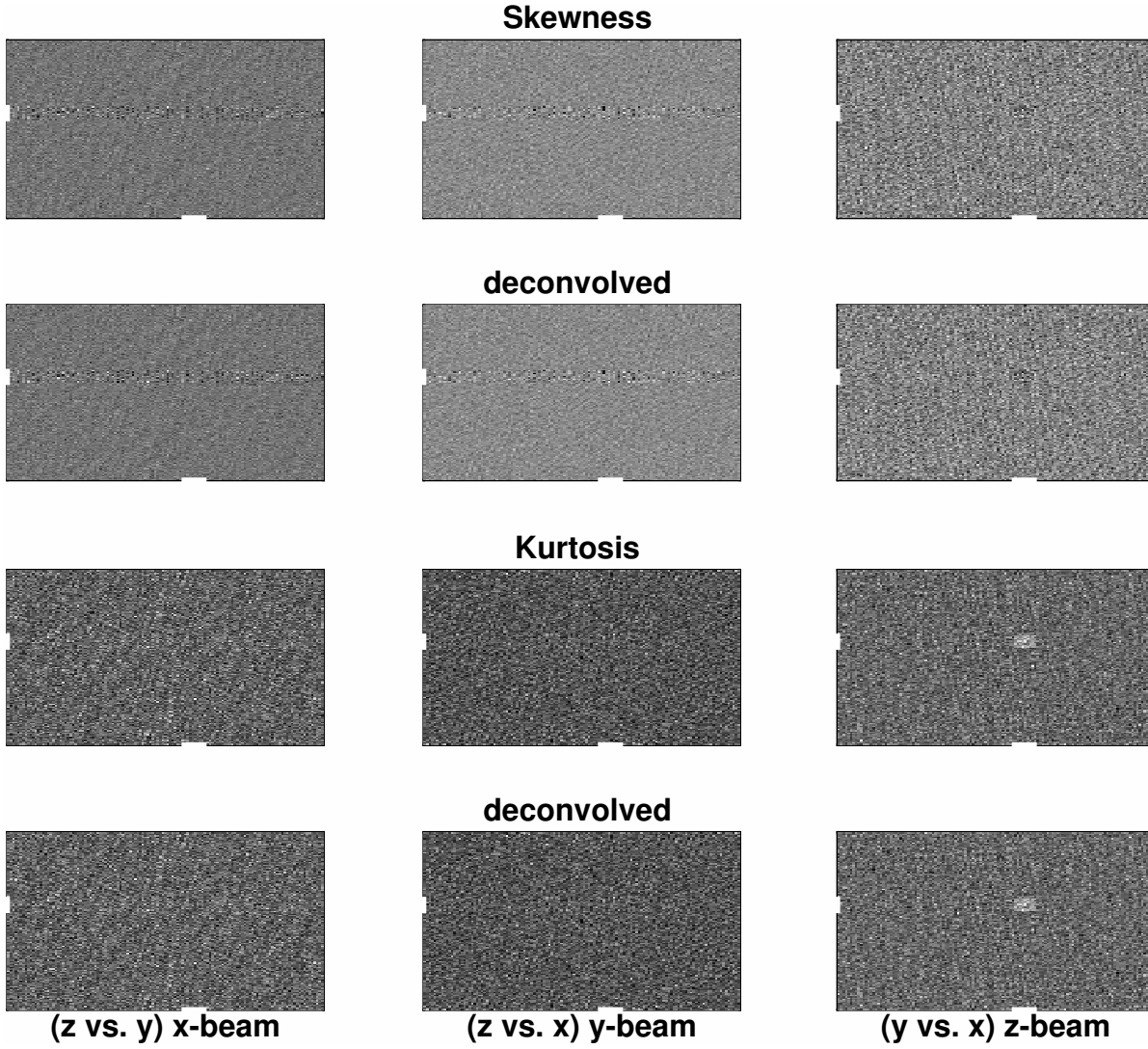


Figure 10. Response to synthetic signal: Skewness and kurtosis statistics maps as in Figure 9, except that a perturbation to the phases on the z-beams was introduced in the range of the 8 spatial x- and y- frequencies indicated by white lines near the bottom and left margins of the plots. Since this perturbation operates along the z-beams, it appears in the plots only for the z-statistic as a function of that in a perpendicular direction – y or x in the first two columns, respectively; x and y in the third. The ranges of the values of the statistics are essentially unchanged by the perturbation, so the grayscale bars are suppressed for clarity.

Figure (10) presents analysis of the skewness and kurtosis of the phase data cube for the Fourier transform of the actual xyz-data, but with an artificial modification imposed on the distribution of the phases along the z-direction. This is meant to be an artificial NG signal injected into the data. In a small interval of spatial frequencies (indices between 72 and 80, inclusive) in the x and y directions the z-beam was modified as follows:

$$\phi_n \rightarrow \phi_n + a\phi_{n+1}, n = 1, 3, 5, \dots \quad (19)$$

For the values shown, $a = 2$ for skewness and $a = 0.4$ for kurtosis, these maps show clear NG effects. If these coefficients are cut in half they become barely detectable, and cut to $\frac{1}{3}$ they essentially disappear. This formula is not meant to be anything like a realistic model of a physically reasonable non-Gaussianity process. It is meant to roughly indicate the level of response of these statistics to a rudimentary, localized perturbation of the phase distribution.

This section developed a visual approach to assessing distributions of statistical parameters in a 3D data cube, and applied it to try to detect departures from the hypothesis of IID Fourier phases. In such displays the eye is famously good at perceiving patterns, but also easily fooled by noise fluctuations. Given the display issues of pixelization, contrast, range, color, non-linear scaling, etc., and the difficulty of rigorous analysis of statistical significance of perceived patterns in this kind of image, a more objective approach is called for, as addressed in the next section.

4.2.3. *Higher Criticism*

The previous section explored a large number of statistical tests for the presence of presumably weak NG effects. [Efron \(2011\)](#) gives an overview of the statistical science dealing with this setting, known as *Large-Scale Inference* (LSI), including its historical development and role in so-called *big data* contexts. We now discuss a key LSI methodology that is ideally suited to our problem.

The method of *Higher Criticism* (HC), perhaps more informatively termed *second-level significance testing*, was introduced by [Donoho and Jin \(2004\)](#) following John Tukey’s parable of the young psychologist. Further technicalities are developed in [Donoho and Jin \(2008, 2015\)](#). [Walther \(2011\)](#) presents comparisons of the effectiveness of various HC and HC-like methods on data of different degrees of sparseness. In the 2D CMB context see e.g. ([Cayón, Jin and Treaster 2005](#)) for applications and ([Jin et al. 2005](#)) for comparison of HC with other methods.

The concept threading HC statistical science is *detection of signals that are both rare and weak* – rare in that significant signals occur in only a small fraction of a large number of available samples; weak in that individual signals may not be detectable on their own. In studying large measurement arrays for the possible presence of weak signals one typically computes a statistical parameter serving to test a specific null hypothesis, usually framed in terms of a known distribution for the observable. Each such test would yield a detection if the null hypothesis is rejected in favor of a different hypothesis corresponding to a different distribution.

Our problem is subtly different from identification of individual measurements likely to contain a signal. HC asks “how likely is it that one or more signals are present,” without necessarily identifying them specifically. In contrast the goal of conventional multiple testing is to decide the truth or falsehood of a hypothesis for each of N measurements, or possibly identification of the one most likely to contain a significant signal. On the other hand the goal of HC is to test the joint null hypothesis that all the nulls are true, against the alternative that some presumably small fraction are false.

From these general remarks it is clear that HC is nearly ideally suited to the search for NG in the context described here. The following details should further clarify this statement. The HC statistic for N measurements can be described as follows:

$$HC_N(\alpha) = \sqrt{N} \frac{(\text{Fraction Significant at level } \alpha) - \alpha}{\sqrt{\alpha(1 - \alpha)}} \quad (20)$$

where α is a statistical significance level. This formula implements a comparison between predictions from the empirical and model probability distributions, in the first and second terms in the numerator respectively. But like most frequentist statistics this form is motivated more through analysis of its behavior under various hypotheses than by specific inferential principles.

There is more than one way to formally define the statistic based on this general idea. We use a slightly simplified version of the one in ([Donoho and Jin 2004](#), first equation in Section 1.2):

$$HC_N^*(\alpha) = \max_{1 \leq i \leq N\alpha} \sqrt{N} \frac{(\frac{i}{N} - p_i)}{\sqrt{p_i(1 - p_i)}} \quad (21)$$

where the p_i are the p -values for the N measurements *sorted in increasing order*. The maximization operation here can be seen as sweeping through p -values – from low to high, along the way accumulating possible evidence of the presence of a signal. That is p_i goes from its smallest values (the most likely to reject the null hypothesis) to larger values (weaker rejection of the null) until this formula indicates that tests later in the sequence can’t provide any further evidence against the null. Other than a slight change of notation, the one simplification of equation (21) is that α actually stands for a critical value obtained by considering a range of values; see [Donoho and Jin \(2004\)](#) for details.

Here we propose a hybrid procedure that combines the promise of skewness and especially excess kurtosis ([Jin et al. 2005](#)) with that of Higher Criticism ([Donoho and Jin 2004](#)). The values of these two statistics, computed along beams in the three different coordinate directions, are taken to be tests of the null hypothesis that phases are IID – identically and independently distributed. The values of skewness and excess kurtosis corresponding to this null hypothesis are shown in Table 2. Departures from these values indicate NG may be present. Throughout HC statistics were computed using Matlab software from the above authors posted at <http://www.stat.cmu.edu/~jiashun/Research/software/HC/Readme.txt>.

For reasons that will be described shortly, we are not able to simply read yes-no decisions from these HC values as is done in the cited references. Rather we have to resort to the relative values. For example, the distribution of the HC statistic in a given setting could be determined from random simulations under the null hypothesis; then the

significance of the value from the actual data can be judged by where it falls in this distribution. An exercise in the same vein is depicted in Figure 11, showing the HC statistic for several subsets of the same data discussed in Figure 9 with NG signals of varying strength injected as in Figure 10. In these four panels, for the usual cases (skewness and kurtosis, original and deconvolved), we plot the HC statistic as a function of the strength of the artificial NG signal injected into the data via the relation in Equation (19). The abscissa is the fraction of the a values used in generating the data reported in Figure (10).

The trends shown in this figure can be interpreted as follows. The strong dependence of the kurtosis-based HC statistic on the amplitude of the injected NG perturbation, compared to the weak dependence for skewness-based one, suggests that kurtosis is the better NG discriminant – cf. the previously mentioned conclusion of Jin et al. (2005). Here we interpret the noisy but flat parts of the curves for small abscissae as an indication that the NG signal is too weak to be detected. The position of the sharp rises in the curves indicate when the NG signal becomes strong enough to be detected via the HC statistic. The actual values of the abscissa where these rises start, at fractions of around 0.8 and 0.5 in the skewness and kurtosis cases, are consistent with the above remarks for Figure 10 about how the NG signatures disappear with decreasing amplitude a . Simulations of this sort could also be used to determine upper limits on NG, as long as the nature of the distortion of the phase distribution is known.

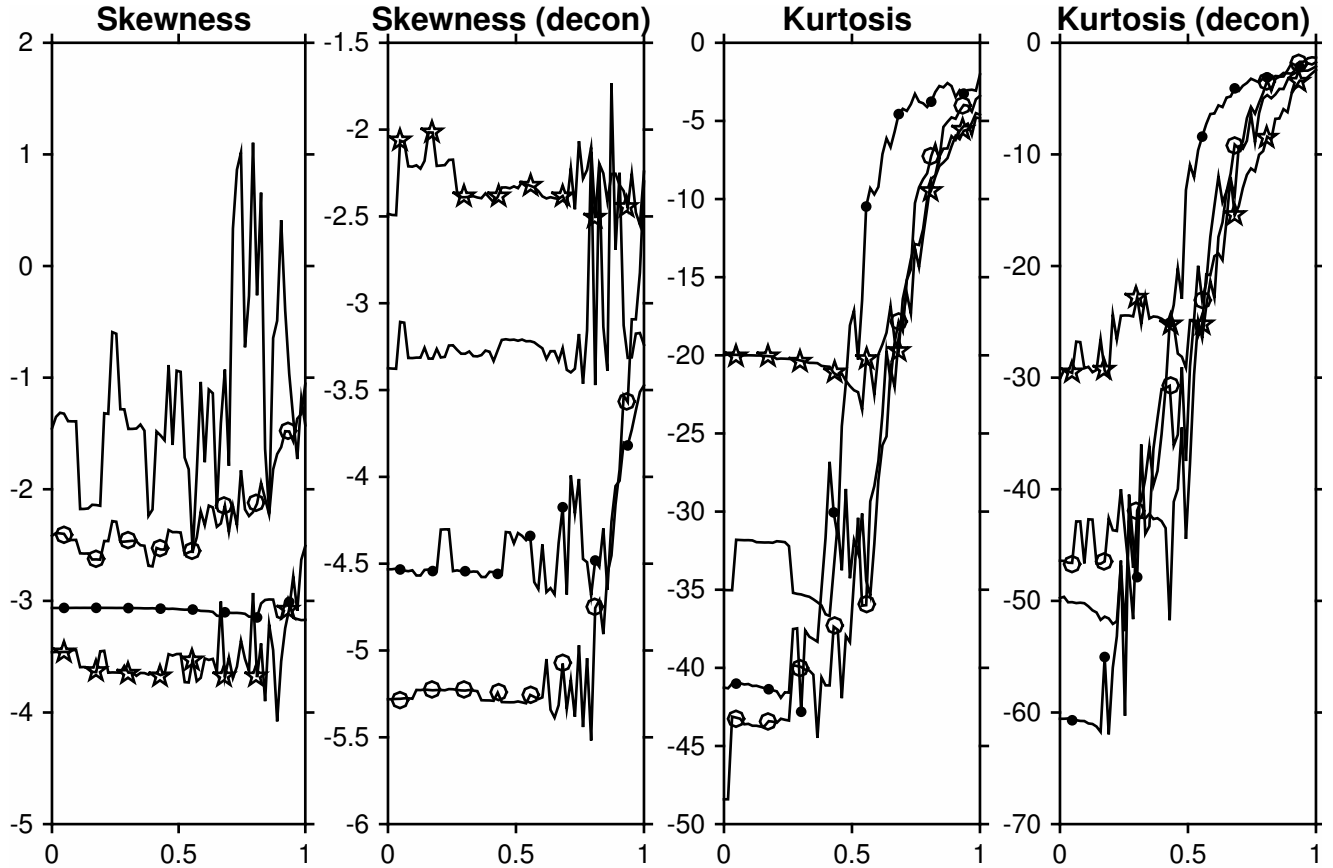


Figure 11. The Higher Criticism statistic. These plots show how the HC statistic behaves as a function of the strength of the synthetic signal. The abscissa gives the fraction of the value of a used in Figure 10: $a = 0$ is the unperturbed case with (presumably) no NG signal present, and $a = 1$ is the full value giving rise to the NG signatures visible in Figure 10. The ordinate is the HC statistic: lines marked with circles, stars, or points are for the x , y , and z beams, respectively; unadorned solid line is for a data set combining all three beams directions. The HC statistics were computed with $\alpha = .004$ and with no lower cutoff of p -values.

Now for the above mentioned caveats regarding the use of HC methodology in the current context. Of course the NG model here is completely arbitrary and simplistic; this problem can be fixed by appeal to predictions of theoretical NG models. The applications developed in the HC literature cited here all use Gaussian null hypotheses – i.e. testing for departures from a normal distribution, and with p -values determined based on this hypothesis. Here the null hypothesis is a very different distribution; the skewness and kurtosis of a uniformly distributed variable are not normal. Absent

an exact expression for the actual distribution, we used an empirical (Monte Carlo) estimate of the distribution function of the HC statistic to generate the uniformly distributed p-values that the HC algorithm expects. This is a key issue, as capture of the true p-value distribution is very important in applications (Emmanuel Candes, private communication). In addition, the HC algorithm we used has two parameters that are not known a priori, and are the subject of some discussion in the literature. We found that our results are not very sensitive to these parameters. We offer the above experiments as suggestive of the promise of HC and related methods, but reliable scientific results will require careful attention to all of these issues.

We close with a few remarks. The problem of assessing a large number of statistical tests is akin to that which faces astronomers who pore over data which, as always, is corrupted by some degree of random noise. After perhaps finding an “interesting” pattern in the data, to assess the statistical significance of the discovery requires correcting for the fact that many cases have been inspected. This procedure is usually carried out by applying a *trials factor* to adjust the post facto statistical significance for the so-called *look elsewhere effect*. In practice, especially when psychological factors are involved, this is difficult or impossible to do post facto. The HC method deals with similar issues but in the somewhat different context described at the beginning of this section. Finally, note that in application HC may in fact yield a large number of statistical results that in turn need to be evaluated in the same way – suggesting terms like *even higher criticism* or *third-level significance testing*. For example HC applied to sets of phase beams of every possible orientation, not necessarily parallel to a coordinate axis, would generate a large number of statistical tests.

4.3. Propagation of the Observational Errors

What we here identify as the only sources of true noise are the fiber collision effects and random measurement errors in the coordinates. Paper II discussed our procedure for mitigating the former. Here we demonstrate that the latter have negligible effects on the results. We simulated 100 realizations of normally distributed heteroscedastic errors (zero mean and standard deviation as given for each galaxy in the data catalog) added to the actual right ascension, declination and redshift values. Power spectra for these data sets were carried out exactly as for the actual data. The relative errors were computed as the standard deviations of the resulting powers divided by the corresponding means. Figure 12 plots these results as functions of spatial frequency. These relative errors are maximum at the highest spatial frequencies, reaching about 0.01 (1%). Overall the effects of coordinate errors are at least several orders of magnitude too small to have any relevance here. **On the other hand, this analysis has ignored systematic errors and the likely possibility of correlated errors induced by systematics. These issues should be addressed in future analysis of new larger samples, where such effects might be more easily assessed.**

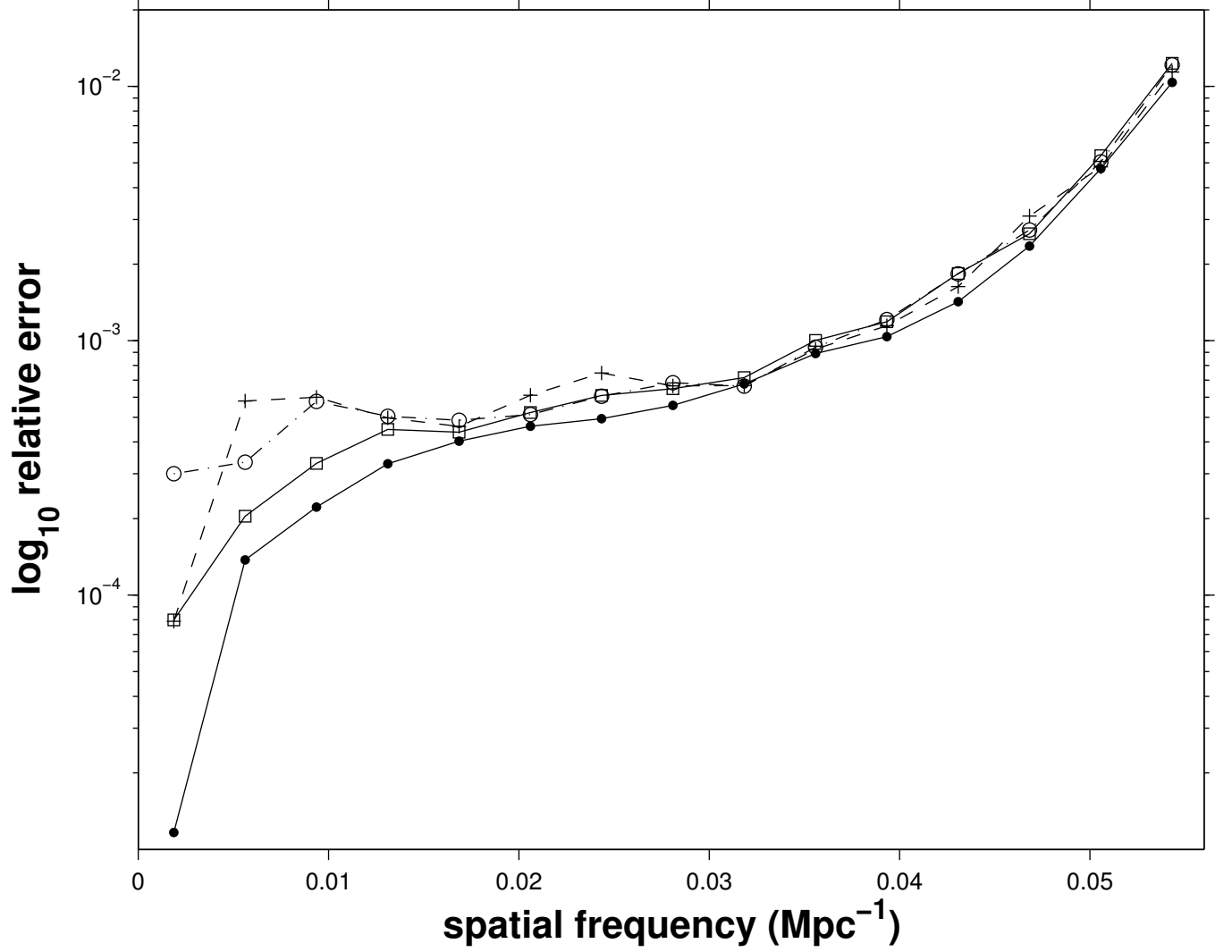


Figure 12. Relative uncertainty from propagation of the observational coordinate errors. The ratio of the standard deviation to the mean of the power spectrum is plotted against spatial frequency. Solid line with dots: rotation averaged over projected directions. Solid (with squares), dashed (with +) and dot-dashed (with circles): x, y, and z shifts, respectively. In all cases abscissa is the average over the rotation angles or shift.

5. EPILOG

Using 3D Fourier transforms of both unbinned and binned rectangular coordinates in the volume limited sample studied in Papers I and II, we have characterized the spatial distribution of galaxies in the nearby ($z \leq 0.12$) Universe on all accessible scales. The results of a truly direct estimator (with no resolution limit other than that inherent to the data) compare well with the transform of the same data binned in small 3D voxels. This complex Fourier transform 3D data cube was used to estimate the power spectrum – both radial as well as projected in three orthogonal directions. However the emphasis has been on useful ways to represent and analyze Fourier phase information, for example in the context of Gaussianity measures. We believe this phase spectrum approach has much to recommend it over the more commonly used multi-point statistics and related methods. Visual display of maps of skewness and kurtosis of the phase distributions along a large number of beams in the 3D Fourier phase data cubes, and a novel analysis based on Higher Criticism do not provide any significant evidence for the non-uniformity of the distribution of phases which could be caused by non-Gaussianity of the spatial distribution of low redshift galaxies. This result, somewhat surprising in view of various effects that might modify Gaussianity of the initial cosmic density perturbations, should be tested with analysis of new larger redshift survey data.

More generally the methods discussed in the paper are meant to suggest techniques for future use in deeper selections of existing redshift surveys and newer larger ones. One simple example is to study isotropy in more detail than was done in Fig. 3 by computing the power spectrum in all possible directions. This extension can probably benefit from complexity reduction techniques in the very similar *beamlet* context (Donoho and Huo 2002). The even larger big-data challenge of phase spectrum analysis ideally requires randomness tests of all possible subsets of the full panoply of estimated phase values. This search for rare and weak NG signatures can be placed in the highly-multiple testing context that is the subject of Higher Criticism. We feel that this and related techniques are promising head-on approaches to this problem. Novel ideas developed in the 2D CMB context, e.g. Kovács, Carron and Szapudi (2013), can probably be generalized to 3D and also applied effectively. Even better would be to develop more tests for NG signatures guided by theory and the results of computational cosmological simulations.

We are grateful to the NASA-Ames Director’s Discretionary Fund and to Joe Bredekamp and the NASA Applied Information Systems Research Program for support and encouragement. We thank the anonymous referee for comments that much improved the manuscript. Thanks goes to Ani Thakar and Maria Nieto-Santisteban for their help with our many SDSS casjobs queries. Michael Blanton’s help with using his SDSS NYU–VAGC catalog is also very much appreciated. We are also grateful to Chris Henze, Roger Blandford, Elliott Bloom, Andrew MacFadyen, Jay Norris, Pratyush Pranav, Aaron Roodman, Alex Silbergleit, Luis Teodoro, Bob Wagoner, Emmanuel Candes, and the referee for helpful suggestions.

Funding for the SDSS has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is <http://www.sdss.org/>.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy, the Max-Planck-Institute for Astrophysics, New Mexico State University, University of Pittsburgh, Princeton University, the United States Naval Observatory, and the University of Washington.

This research has made use of NASA’s Astrophysics Data System Bibliographic Services.

REFERENCES

- AAde, P. et al. 2014, *Astronomy and Astrophysics*, 571, A23
- Alam, S., Ata, M., Bailey, S. et al. 2016, The clustering of galaxies in the completed SDSS-III Baryon Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample, arXiv: 1607.03155
- Aschwanden, M. 2011, *Self-Organized Criticality in Astrophysics: The Statistics of Nonlinear Processes in the Universe*, Springer: Heidelberg.
- Bardeen, J., Bond, J., Kaiser, N., and Szalay, A. 1996, *ApJ*, 304, 15.
- Bracewell, R. 1999, *The Fourier Transform and Its Applications*, 2nd Revised Edition, McGraw-Hill, Inc., New York
- Cayón, L., Jin, J. and Treaster, A. 2005, *MNRAS* 362, 826
- Carron, J. 2011, *ApJ* 738, 86
- Carron, J. and Szapudi, I. 2015, What does the N-point function hierarchy of the cosmological matter density field really measure ?, arXiv:1508.04838
- Carron, J., Wolk, M. and Szapudi, I. 2015, *MNRAS* 453, 450
- Chiang, L., Naselsky, P., Verkhodanov, O. and Way, M. 2003 *ApJ*, 590, L65
- Chiang, L. and Naselsky, P. 2007, *MNRAS*, 380, L71
- Chiang, L. and Coles, P. 2000, *MNRAS*, 311, 809
- Chiang, L., Naselsky, P., and Coles, P., 2004, *ApJ*, 602, L1
- Coles, P. and Chiang, L. 2000, *Nature*, 406, 376
- Coles, P. 2000, Large-Scale Structure, Theory and Statistics, p. 593, in *Phase transitions in the early universe: Theory and observations. Proceedings, NATO ASI, International School of Astrophysics 'Daniel Chalonge', 8th Course, de Vega, H. J., Khalatnikov, I. M. and Sanchez, N. G. eds., Dordrecht, Netherlands: Kluwer Academic*, arXiv: 0103017
- Coles, S., Percival, W. et al 2005, *MNRAS*, 362, 505
- Contaldi, C., Ferreira, P., Magueijo, J. and Górski, K. 2000. *ApJ*, 534, 25
- Donoho, D., and Huo, X. 2002, "Beamlets and multiscale image analysis." *Multiscale and multiresolution methods*, Vol. 20, Springer Berlin Heidelberg, 2002. 149-196. *Multiscale and Multiresolution Methods*, Volume 20 of the series *Lecture Notes in Computational Science and Engineering* pp 149-196
- Donoho, D., and Jin, J. 2004, *Annals of Statistics* 32.3, 962
- Donoho, D., and Jin, J. 2008, *Proceedings of the National Academy of Sciences of the United States of America*, 105, 14790
- Donoho, D., and Jin, J. 2015, for *Rare and Weak Effects Statistical Science*, 30, 1
- Dore, O. et al. 2015, *Cosmology with the SPHEREX All-Sky Spectral Survey*, arxiv:1412.4872
- Efron, B., 2011, *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge University Press, Cambridge. MR2724758
http://statweb.stanford.edu/~ckirby/brad/LSI/monograph_CUP.pdf
- Eggemeier, A., Battefeld, T., Smith, R., Niemeyer, J. 2015, *MNRAS*, submitted.
- Efstathiou, G. and S.J. Moody, S. 2001, *MNRAS* 325, 1603
- Feldman, H., Kaiser, N., and Peacock, J. 1994, *ApJ*, 426, 23
- Ferreira, P. and Jagueijo, J. 1997, *Phys. Rev. D* 55, 3358, arxiv: 9610174
- Gunn, J. A *Mathematical Framework for Discussing the Statistical Distribution of Galaxies in Space and its Cosmological Implications*, Ph. D. Thesis, CalTech, 1965.
- Hikage, C., Matsubara, T., Suto, Y., Park, C., Szalay, A., and Brinkmann, J. 2005, *PASJ*, 57, 709
- Hikage, C., Komatsu, E., and Matsubara, T. 2006, *ApJ*, 653, 11
- Hikage, C., Coles, P., Groiissi, M. et al. 2008, *MNRAS*, 385, 161
- Jin, J., Starck, J.-L., Donoho, D., Aghanim, N., and Formi, O., 2011, *Annals of Statistics*, 39, 2533
- Kitaura, F. 2010, *MNRAS*, 420, 2737
- Kovács, A., Szapudi I., and Frei, Z. 2013, *Astronomische Nachrichten*, 334, 1020 arXiv: 1308.0837
- Kovács, A., Carron, J. and Szapudi, I. 2013 *MNRAS* 436, ?
- Landy, S. and Szalay, A. 1993, *ApJ*, 412, 64L
- Lentati, L., Hobson, M., and Alexander, P. 2014, *MNRAS*, 444, 3863
- Limber, N. 1953, *The Analysis of Counts of the Extragalactic Nebulae in Terms of a Fluctuating Density Field*, *ApJ*, 117, 134L
- Mannell, R.H., 1990, *Proc. Third Australian International Conference on Speech Science and Technology*, Melbourne
- Martínez-González, E. 2009, in *Data Analysis in Cosmology*, *Lecture Notes in Physics*, vol. 665. Edited by V. J. Martnez, E. Saar, E. Martnez-Gonzlez, and M.-J. Pons-Bordera. Berlin: Springer, 2009., p.79-120
- Matsubara, T. 2007, *ApJS*, 170, 1
- Nadelsky, P., Chiang, L., Olesen, P. and Novikov, I. 2005, *PhRvD*, 72, 063512
- Nadelsky, P., Doroshkevich, A., and Verkhodanov, O. 2003, *ApJ*, 599, L53
- Nadelsky, P., Doroshkevich, A., and Verkhodanov, O. 2004, *MNRAS*, 349, 695
- Oppenheim, A. and J. S. Lim, J. 1981, *Proceedings of the IEEE*, 69, 529
- Peebles, J. 1975, *ApJ*, 185, 413
- Peebles, J. and Hauser, M. 1974, *ApJ*, 185, 757
- Peebles, J. and Hauser, M. 1974, *ApJS*, No. 28, 19
- Percival, W., Verde, L., and Peacock, J. 2004, *MNRAS*, 347, 645
- Percival, W., Nichol, R., Eisenstein, D., Frieman, J., Fukugita, M., Loveday, J., Pope, A., Schneider, D., Szalay, A., Tegmark, M., Vogeley, M., Weinberg, D., Zehavi, I., Bahcall, N., Brinkmann, J., Connolly, A., and Meiksin, A. 2007, *ApJ*, 657, 645
- Querre, P, Starck, J.-L., and Martínez, J. *SPIE*, 4847
- Raccanelli, A., Monanari, F., Bertacca, D., Dore, O., and Durrer, R. 2015, arXiv: 1505.06179v1
- Rocha, G., Magueijo, J., Hobson, M., and Lasenby, A. 2001, *Phys. Rev. D* 64, 063512
- Sánchez, A. and Cole, S. 2008, *MNRAS*, 385, 830
- Scargle, J. 1981, *ApJS*, 45, 1
- Scargle, J. 1990, *ApJ*, 359, 469.
- Sefusatti, E. and Komatsu, E. 2007, *Phys. Rev. D* 76, 083004
- Slepian, Z. and Eisenstein, D. 2015, *MNRAS* 448, 1, 9 arxiv: 1411.4052
- Slepian, Z. and Eisenstein, D. 2015, arxiv: 1506.02040
- Slepian, Z. and Eisenstein, D. 2015, arxiv: 1506.04746
- Strauss, M. A., et al., 2002 *AJ*, 124, 1810
- Tegmark, M. et al. 1998, error in bibliography of the other Tegmark paper?
- Tegmark M., Hamilton A., Xu Y., 2002, *MNRAS*, 335, 887
- Tegmark, M., Blanton, M. et al. 2004, *ApJ*, 606, 702
- Vogeley M. and Szalay A. 1996. *ApJ*, 465, 34
- Walther, G. 2013, in *Probability to Statistics and Back: High-Dimensional Models and Processes - A Festschrift in Honor of Jon A. Wellner*. M. Bannerjee, F. Bunea, J. Huang, V. Koltchinskii, M.H. Maathuis (eds.), *Inst. Math. Statistics*, 317-326
arXiv:1111.0328 [stat.ME].
- Way, M.J., Gazis, P.R. & Scargle, J.S. 2011, *ApJ*, 727, 48
- Way, M.J., Gazis, P.R. & Scargle, J.S. 2015, *ApJ*, 799, 95
- Wolk, M., Carron, J., and Szapudi, I. 2015, *MNRAS* 454, ?
- Wolstenhulme, R., Bonvin, C., and Obreschkow, D. 2015, *ApJ*, 804, 132
- Yu, J. & Peebles, J. 1969, *ApJ*, 158, 103

APPENDIX

A. CHECKING THE FORMALISM USING THE INVERSE FOURIER TRANSFORM

It is useful to check how well our Fourier transform estimates capture the information in the galaxy coordinate data. The discrete Fourier transform of evenly spaced voxels is exactly invertible and therefore lossless, and so is the direct transform in Eqs. (4) and (5) in the limit of an infinite number of frequencies. Nevertheless it is of some interest to see how this limit is approached by comparing its inverse transform against the raw data. For this limited purpose a rough visual check suffices, since a precise goodness-of-fit metric, involving comparison of an effectively continuous representation with point data, is difficult.

In the direct transform there is no binning of galaxy positions, so if the Fourier transform were to be evaluated at an infinite number of spatial frequencies the inverse transform would exactly reproduce the data. That is the function

$$F_x(\mathbf{x}) = \int \mathbf{F}(\mathbf{k}) e^{-i\mathbf{k} \cdot \mathbf{x}} d\mathbf{k} \quad (\text{A1})$$

would vanish except for unit delta functions at each of the galaxy positions. Normalization is not important here, so the factor $\frac{1}{(2\pi)^{3/2}}$ sometimes written in front of the right-hand side of this equation is omitted. Inserting Eq. (4) into this expression we have

$$F_x(\mathbf{x}) = \sum_{\mathbf{n}=1}^N \int e^{-i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}_n)} d\mathbf{k} \quad (\text{A2})$$

It is well known that this integral is equivalent to a δ -function at x_n , yielding the desired result.

It is useful to investigate how well these exact results apply to necessarily finite numerical computations. To do so we evaluate the expression

$$f(x, y, z) = \sum_{k_x, k_y, k_z} F(k_x, k_y, k_z) e^{-i(k_x x + k_y y + k_z z)} \quad (\text{A3})$$

appropriately normalized, against the raw data in Figure A1. The total number of spatial frequencies increases as the third power of the number of frequencies in each dimension, but it is nevertheless feasible to use a frequency grid that well resolves the relevant spatial structure in all three directions. The spatial frequencies were taken to be the usual integer multiples of the fundamental frequency of $\frac{1}{1082} \text{Mpc}^{-1}$. This denominator is approximately twice the maximum of the x, y and z ranges of the data, to eliminate wraparound.

Since the forward transform can be evaluated at any set of spatial frequencies, it is expedient to use an FFT algorithm⁷ to evaluate the expression above. We reconstructed 3D data points at every location in the $N_k \times N_f \times N_k$ array (N_k^3 voxels) where the value of $f(x, y, z)$ in equation (A3) exceeds a threshold. This threshold value was chosen to yield the same number of points (139,798) as in the raw data. The figure shows xy-projections of the points contained in a 12.5 Mpc thick slice in the z-coordinate, isolating roughly 5,000 points in all three panels. The limited frequency range dictates that the reproductions are smoothed representations of the galaxy data. The sequence in this figure demonstrates that increasing the number of spatial frequencies reproduces the discrete raw data with improved accuracy. Note that panel (c) with $N_k = 256 + 1$ seems to have more points than (d) for $N_k = 512 + 1$; in fact both have the same number, those in (d) more closely following the narrow filaments and other structures (with a consequent increase in overplotting of points) and therefore more faithfully reproducing the data. The key point is that information about the discrete structure at a broad range of scales, limited only by the resolution of the computation, is contained in the Fourier transform in eq. (5).

⁷ Our expression for the forward transform does not automatically impose the complex conjugate symmetry necessary for the inverse transform computed in this way to be real. To deal with this problem we simply evaluate the forward transform at an odd number of points: one corresponding to zero frequency, $(N - 1)/2$ at positive frequencies, and the remaining $(N - 1)/2$ at the corresponding negative frequencies. This symmetry yields a positive result. Accordingly values for the number of frequencies are written in the form $N + 1$ throughout, where N is even.

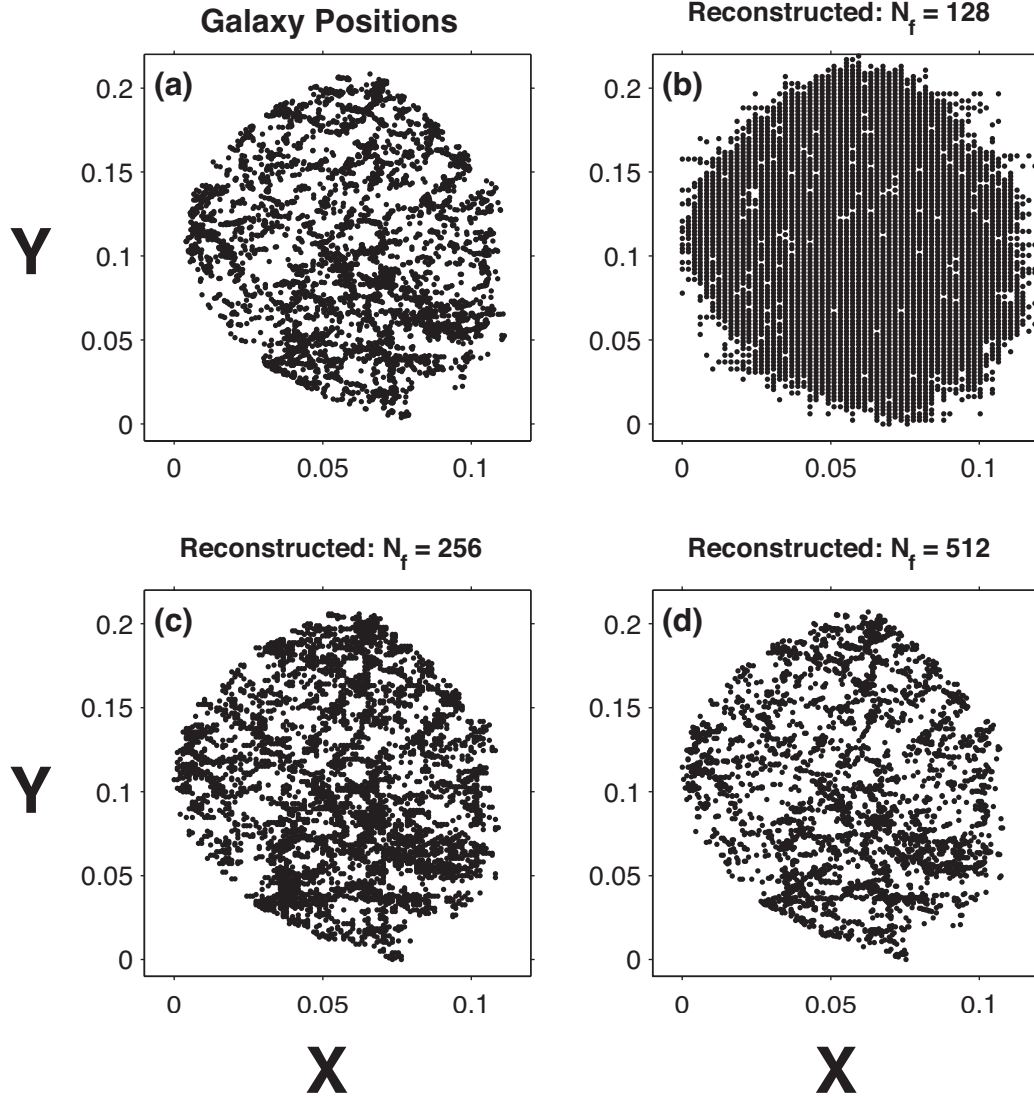


Figure A1. Comparison of x-y projections of thin (12.5 Mpc.) z-slices for (a) the galaxy data; and the corresponding reconstruction with the direct Fourier Transform in Eq. (A3) using (b) 128 frequencies; (c) 256 frequencies, and (d) 512 frequencies. Coordinates are in redshift units (rsu). The effective resolutions of the reconstructions are 16.9, 8.5 and 4.2 Mpc, respectively. Plots of other projections are very similar.

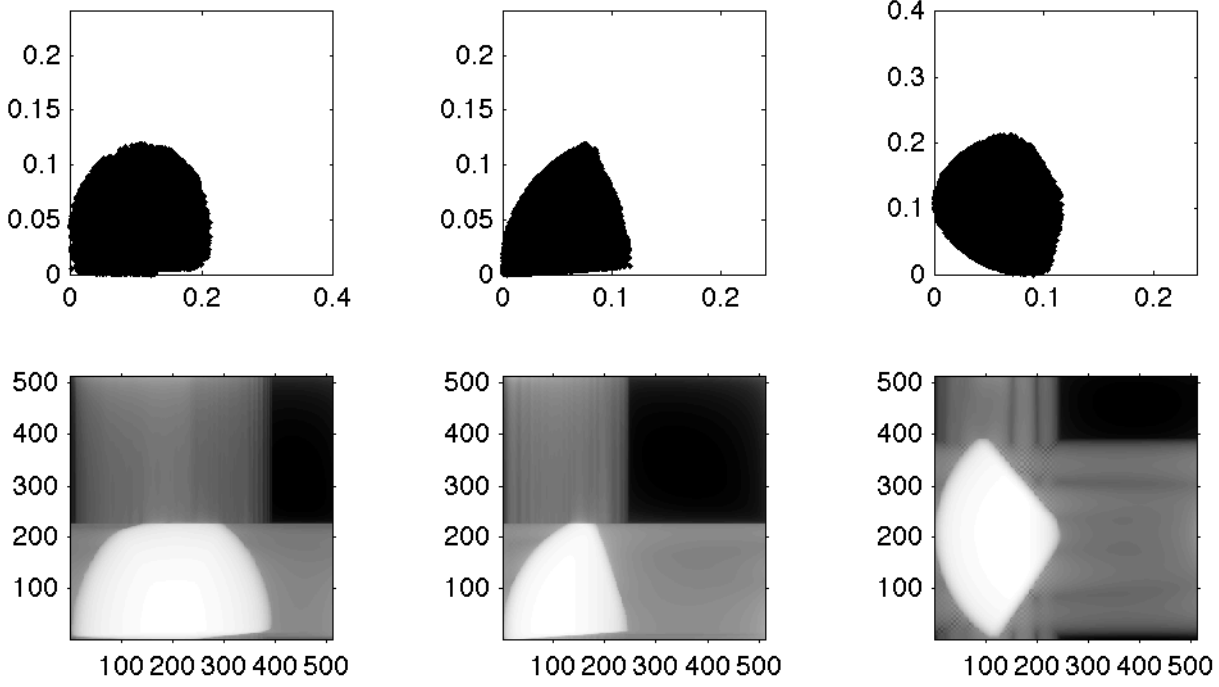


Figure A2. Reconstruction of the selection function. Top row: projections of the raw galaxy coordinates along the x -, y -, and z -axes respectively (empty white spaces indicate the extended ranges employed in the Fourier transforms) for comparison with the corresponding reconstructions via the inverse Fourier transform in the bottom row. These are 2D projections of thin slices oriented as indicated by the axis labels, with the gray-scale representing the reconstructed density. Here the grid size is .001 redshift units, for which the relative volume error in the cuboid representation is about one part in ten thousand.

In the same way the projected density plots in Figure A2 demonstrate that the inverse transform of the selection function Fourier transform is essentially a uniform solid corresponding to the observational data space. This procedure accounts for incorporating only galaxies inside the window, but of course does not in any way replace or estimate data outside of the window.

B. MATLAB CODE FOR 3D FOURIER TRANSFORMS

The MatLab code shown here computes the Fourier transform of the window function. The first part finds a refined partition of the actual data space – here taken to be the convex hull of the galaxy positions – into cuboids with x and y coordinates in an evenly spaced rectangular grid. The z coordinates of the end faces of the cuboid are taken to be those of the two points where a infinite line, parallel to the z -axis and passing through the center of the cuboid in the $x - y$ plane, intersects faces of the convex hull. It is easy to see that each such line must intersect either 0 or 2 such faces, the former making no contribution. Linearity of the transform then enables evaluation of the transform by simply summing the transforms of the cuboids.

```
function data_out = window_ft( data_in )
% Compute Fourier transform of 3D window ("selection") function
    xyz = data_in.xyz;
    size_grid = data_in.size_grid;
    kk_vec = data_in.kk_vec;

    nn_kk = length( kk_vec );
    phi_vec = i * ( 1 - exp( i * size_grid * kk_vec ) ) ./ kk_vec;
    id_freq_zero = find( kk_vec == 0 );
    phi_vec( id_freq_zero ) = size_grid;% Limiting value

% Set up xyz grid

xyz_max = max( xyz );
```



```

xyz_min = min( xyz );
xyz_range = xyz_max - xyz_min;

nn_grid = ceil( xyz_range / size_grid );
dist_excess = size_grid * nn_grid - xyz_range;
xyz_start = xyz_min - 0.5 * dist_excess;
xyz_stop = xyz_max + 0.5 * dist_excess;

xx_grid_vec = xyz_start(1):size_grid:xyz_stop(1);
yy_grid_vec = xyz_start(2):size_grid:xyz_stop(2);

nn_xx = length( xx_grid_vec );
nn_yy = length( yy_grid_vec );
zz_min_mat = data_in.zz_min_mat;
zz_max_mat = data_in.zz_max_mat;

% Identify the 3D facets of the hull

[ kkk, volume_hull ] = convhulln( xyz );
[ num_hull_facets, dum_3 ] = size( kkk );

hull_facets_3d_area = zeros( num_hull_facets, 1 );
hull_facets_2d_area = zeros( num_hull_facets, 1 );% area of xy projection

facet_xx_min = zeros( num_hull_facets, 1 );
facet_xx_max = zeros( num_hull_facets, 1 );

facet_yy_min = zeros( num_hull_facets, 1 );
facet_yy_max = zeros( num_hull_facets, 1 );

for ii_hull_facet = 1: num_hull_facets

    id_this = kkk( ii_hull_facet, : );% Indices of the 3 points of this facet

    xx_this = xyz( id_this, 1 );% Coordinates of triangle vertices
    yy_this = xyz( id_this, 2 );
    zz_this = xyz( id_this, 3 );

    [ kk_2d, hull_facets_2d_area( ii_hull_facet ) ] = ...
        convhulln( [ xx_this yy_this ] );

    x = xx_this(:)';
    y = yy_this(:)';
    z = zz_this(:)';
    ons = [1 1 1];
    hull_facets_3d_area( ii_hull_facet ) = ...
        0.5*sqrt(det([x;y;ons])^2 + det([y;z;ons])^2 + det([z;x;ons])^2);

    facet_xx_min( ii_hull_facet ) = min( xx_this );
    facet_xx_max( ii_hull_facet ) = max( xx_this );

    facet_yy_min( ii_hull_facet ) = min( yy_this );
    facet_yy_max( ii_hull_facet ) = max( yy_this );

```

```

end

data_out.hull_facets_2d_area = hull_facets_2d_area;
data_out.hull_facets_3d_area = hull_facets_3d_area;

zz_min_mat = zeros( nn_xx - 1, nn_yy - 1 );
zz_max_mat = zeros( nn_xx - 1, nn_yy - 1 );

% Go through all of the 2D pixels (in the xy-plane)
for ii_xx = 1: nn_xx - 1
    % X center of the pixel
    xx_mid_pix = ( xx_grid_vec( ii_xx ) + xx_grid_vec( ii_xx + 1 ) ) / 2;

    %=====

    for ii_yy = 1:nn_yy - 1

        % Y center of the pixel
        yy_mid_pix = ( yy_grid_vec( ii_yy ) + yy_grid_vec( ii_yy + 1 ) ) / 2;

        % Only test facets that overlap this pixel
        id_hull_good = find( facet_xx_min <= xx_mid_pix & ...
                             facet_xx_max >= xx_mid_pix & ...
                             facet_yy_min <= yy_mid_pix & ...
                             facet_yy_max >= yy_mid_pix );

        num_good_this = length( id_hull_good );

        if ~isempty( id_hull_good )

            area_test = hull_facets_2d_area( id_hull_good );
            area_aug_vec = Inf * ones( size( area_test ) );
            xx_good_mat = zeros( num_good_this, 3 );
            yy_good_mat = zeros( num_good_this, 3 );
            zz_good_mat = zeros( num_good_this, 3 );

            % test each overlapping triangle
            for ii_run = 1: num_good_this

                id_hull_facet = id_hull_good( ii_run );

                id_this = kkk( id_hull_facet, : ); %Indices of the 3 points
                xyz_1 = xyz( id_this, 1 ); % xyz of triangle vertices
                xyz_2 = xyz( id_this, 2 );
                xyz_3 = xyz( id_this, 3 );

                xx_good_mat( ii_run, : ) = xyz_1;
                yy_good_mat( ii_run, : ) = xyz_2;
                zz_good_mat( ii_run, : ) = xyz( id_this, 3 );

                xyz_1a = [ xyz_1' xx_mid_pix ]'; % Add pixel center
                xyz_2a = [ xyz_2' yy_mid_pix ]';

                % Compute area of this facet augmented by the pixel center

```

```

        [ kk_2d, area_aug_vec( ii_run ) ] = convhulln( [ xyz_1a xyz_2a ] );

    end

    % If the augmented area is not changed (does not increase)
    % the pixel lies in xy projection of (2) hull facets
    index_good = find( area_aug_vec <= area_test );

    % Now work with these (two) facets

    if ~isempty( index_good )

        id_facets = id_hull_good( index_good );% index of facets
        num_facets = length( id_facets );

        %-----
        % To estimate z range for this pixel:
        %   find the z-coordinates of the 2 points where
        %   a line through the pixel center, parallel to the
        %   z-axis, intersects the hull facets
        %-----

        zz_pix_vec = zeros( 2, 1 );

        for ii_facet = 1: num_facets

            ii_hull_facet = index_good( ii_facet );
            xyz_1 = xx_good_mat( ii_hull_facet, : )';
            xyz_2 = yy_good_mat( ii_hull_facet, : )';
            xyz_3 = zz_good_mat( ii_hull_facet, : )';

            aa_mat = [ xyz_1 xyz_2 ones( 3,1 ) ]; % projected coordinates

            [ abc, rrr ] = linsolve( aa_mat, xyz_3 );

            err_vec = aa_mat * abc - xyz_3;
            if max( abs( err_vec ) > 1.e-12 )
                error('error in linear solve')
            end

            aa_pix = [ xx_mid_pix yy_mid_pix 1 ];
            zz_pix = aa_pix * abc;
            zz_pix_vec( ii_facet ) = zz_pix;

        end

        zz_min_mat( ii_xx, ii_yy ) = min( zz_pix_vec );
        zz_max_mat( ii_xx, ii_yy ) = max( zz_pix_vec );

    end

else

end

```

```

end

end

data_out.zz_min_mat = zz_min_mat;
data_out.zz_max_mat = zz_max_mat;

% Compute the error: difference between the volume of the convex hull
% and the sum of the volumes of the cuboids identified as contained
% in the convex hull. In the limit of grid_size --> 0 this is a way to
% define the integral giving the volume of the convex hull
volume_pix = ( size_grid .^ 2 ) * sum( sum( zz_max_mat - zz_min_mat ) );
volume_error = abs( volume_hull - volume_pix )/ volume_hull;
data_out.volume_error = volume_error;

end

%=====
%           compute sum exp( i dx * dk ) (vector dot product)
%           3D vector range and spatial frequency
%=====

window_fourier_transform = zeros( nn_kk, nn_kk, nn_kk );

for ii_xx = 1: nn_xx - 1

    xx_this = xx_grid_vec( ii_xx );
    wft_xx = phi_vec .* exp( i * xx_this * kk_vec );

    for ii_yy = 1: nn_yy - 1

        yy_this = yy_grid_vec( ii_yy );
        wft_yy = phi_vec .* exp( i * yy_this * kk_vec );

        zz_min = zz_min_mat( ii_xx, ii_yy );
        zz_max = zz_max_mat( ii_xx, ii_yy );

        if zz_min ~= 0

            wft_zz = i * ( exp( i * zz_max * kk_vec ) - ...
                           exp( i * zz_min * kk_vec ) ) ./ kk_vec;
            wft_zz( id_freq_zero ) = zz_max - zz_min;

            for ii_ff_xx = 1: nn_kk
            for ii_ff_yy = 1: nn_kk
            for ii_ff_zz = 1: nn_kk

                wft( ii_ff_xx, ii_ff_yy, ii_ff_zz ) = ...
                    wft_xx( ii_ff_xx ) .* ...
                    wft_yy( ii_ff_yy ) .* ...
                    wft_zz( ii_ff_zz );

            end

        end

    end

end

```

```
        end
    end

    window_fourier_transform = window_fourier_transform + wft;
end
end
end
data_out.window_fourier_transform = window_fourier_transform;
data_out.zz_min_mat = zz_min_mat;
data_out.zz_max_mat = zz_max_mat;
```